Commissariat à la protection de la vie privée du Canada

Office of the Privacy Commissioner of Canada

Désignation sécuritaire / Security Classification	Pages totales / Total pages (non compris les pièces jointes / not including attachments)
UNCLASSIFIED	12
	✣ Nombre de pièces jointes / Number of attachments: 3
No. de suivi / Tracking No.: CTS-096133	No. de document / Document No.: 7777-6-489504

NOTE DE BREFFAGE **BRIEFING NOTE**

Titre / Title

Entretien avec le Dr. Khaled El Emam
sur la réglementation des données synthétiques le 14 février 2022 /
Interview with Dr. Khaled El Emam
on the regulation of synthetic data on February 14, 2022

OBJET / PURPOSE: Pour information / for information

ENJEU / ISSUE: Dr. Khaled El Emam has requested a 45-minute interview with the Commissioner to discuss the Commissioner's views on the regulation of synthetic data within Canada. The interview is part of a research project for which Dr. El Emam received funding through the OPC's contributions program.

APERÇU / OVERVIEW: Dr. El Emam's full research project is entitled "A Pan-Canadian Descriptive Study of Privacy Risks from Data Synthesis Practices within the Evolving Canadian Legislative Landscape." It has three main research phases. The interview he has requested with the Commissioner is part of the third phase. It is entitled "Interview Study of Canadian Privacy Regulators on Regulating Synthetic Data."

The other two phases are "An overview of data synthesis (environmental scan / literature review)" and "A legal analysis of data synthesis under Part I of PIPEDA and the *Consumer Privacy Protection Act* (CPPA) component of the *Digital Charter Implementation Act, 2020* (Bill C-11)."

Dr. El Emam's research project describes the purpose of the third phase / interview study as follows:

> The objective of this third phase of the project is to understand the perspectives of Canadian regulators on the same four areas outlined in the legal review: i) Identifying privacy risks associated with the use of synthetic data as a PET; ii) Assessing the risks against the current legal framework (PIPEDA and the proposed provisions of the CPPA); iii) Identifying areas where gaps are perceived in the current legal framework; and (iv) Recommendations on how these gaps may be addressed. We also seek to gather information on regulators' experiences with

> synthetic data within their jurisdictions (such as case studies, complaints, investigations, and queries that have come through their offices).[1]

The Commissioner is one of the Canadian privacy regulators Dr. El Emam is interviewing. Dr. El Emam also plans to interview representatives from the offices of the provincial and territorial privacy commissioners as well as members of the privacy and access law section of the Canadian Bar Association.[2]

The interviews will be compiled into a summarized report that will be published. Dr. El Emam states that, "The report and any subsequent publications will not attribute any comment to an identifiable interviewee. All reported information will be of trends and aggregate information."[3]

In advance of the interview, Dr. El Emam has sent background materials and a list of nine interview questions. The background materials consist of a short 1-page preamble to the interview questions as well as an excerpt from Dr. El Emam's recently published book, *Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data*.[4] The excerpt is a legal analysis of three non-Canadian privacy laws[5] and their impact on synthetic data generation. It was not written by Dr. El Emam (or one of the book's listed co-authors). Mike Hintze from the US-based law firm Hintze Law wrote the section.

The interview is scheduled to take place on February 14, 2022 from 11:00–11:45. It will be done remotely over Zoom. David Weinkauf will accompany the Commissioner in the interview.

The nine interview questions are intended to guide, but not constrain, the discussion. Dr. El Emam notes that, "We will deviate from this list if the conversation leads us in new directions."[6]

The interview will be recorded and transcribed to ensure important points are not missed. The recordings will be deleted after the study. The transcript will be kept for seven years in case questions arise about publications from the work.

Dr. El Emam will provide the OPC with a copy of the final report upon submission to the contributions program. However, he is also willing to allow us to review a draft, but notes that he expects turn-around times for the review cycle to be short.

---

[1] El Emam, "A Pan-Canadian Descriptive Study of Privacy Risks from Data Synthesis Practices within the Evolving Canadian Legislative Landscape," p. 20.
[2] Ibid., p. 21.
[3] El Emam, Interview background materials.
[4] Khaled El Emam, Lucy Mosquera and Richard Hoptroff, *Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data* (Sebastopol, California: O'Reilly, 2020).
[5] Specifically, the laws are the EU *General Data Protection Regulation* (GDPR), the U.S. *Health Insurance Portability and Accountability Act* (HIPAA) and the *California Consumer Privacy Act* (CCPA).
[6] El Emam, Interview background materials.

CONTEXTE / BACKGROUND: Synthetic data is an old de-identification technique that has recently undergone notable advancements in terms of its functionality and scope of application. Versions of it have existed since at least the 1990s. However, the recent use of artificial intelligence (AI) / machine learning (ML) statistical modelling techniques, combined with greater computing power, have increased its ability to capture further and more nuanced relationships among data points, while protecting the identity of individuals within the original source dataset.

These advancements have made synthetic data an attractive de-identification technique to apply to "big data." Using AI/ML techniques, synthetic data is able to de-identify large sets of training data, which, in turn, can be used to power the development of further AI/ML systems.

Given this role as an enabler of AI/ML systems, synthetic data is currently receiving a lot of attention. Forrester has named it one of five "key advances" to realizing the next level of AI for businesses.[7] Gartner has predicted that, "By 2024, 60% of the data used for the development of AI and analytics projects will be synthetically generated."[8]

What is synthetic data?

Essentially, synthetic data is fake data that retains the same or similar statistical properties as the source data from which it is generated, but with enough individual-level differences to protect the identity of individuals. In general, there are four components to be aware of:

- *The source data.* This is the dataset whose statistical properties the synthetic data is trying to emulate. Other than the removal of variables with no analytic utility, i.e., variables that are deemed not useful to potential secondary analyses, it does not undergo any data transformations. This means that, if it is about individuals, it will likely contain personal information. It will almost certainly contain quasi-identifiers (e.g., age, gender, race, etc.) but may even contain some direct identifiers (e.g., facial image, address, DNA, etc.) depending on the scope of future analyses.

- *The generative model*. This is the statistical model used to generate the synthetic data. It is derived from the source data. There are multiple methods by which to derive a generative model. However, the one that is garnering the most attention today is the use of AI/ML tools. Using AI/ML tools, a generative model is able to "learn" the statistical properties of the source data without making explicit assumptions about the underlying distributions of variables and correlations among them. This enables it to discover and replicate more relationships in the source data, especially in the case of high-dimensional datasets.

- *The synthetic data*. This is the data generated from the generative model. If done properly, it will capture the statistical properties of the source data without

[7] Forrester, "Five Key Advances Will Upgrade AI To Version 2.0 For Enterprises," https://www.forrester.com/press-newsroom/five-key-advances-will-upgrade-ai-to-version-2-0-for-enterprises/.
[8] https://blogs.gartner.com/andrew_white/2021/07/24/by-2024-60-of-the-data-used-for-the-development-of-ai-and-analytics-projects-will-be-synthetically-generated/.

replicating individual-level records. Typically, it is generated by taking samples of data points drawn from the distribution of the generative model.

- *Privacy and utility metrics*. These measure the similarities and differences between the source data and the synthetic data. There are multiple metrics available. If the source data and synthetic data are too similar, individuals may be re-identified. However, if they are too different, the synthetic data will be inaccurate, potentially leading to downstream harms for individuals who are subject to AI/ML tools trained on the synthetic data.

In addition, there are two general types of synthetic data:

- *Fully synthetic data*. This is where the full set of variables in the source data are synthetically generated.

- *Partially synthetic data*. This is where only the quasi-identifiers or other sensitive variables in the source data are synthetically generated. The remaining variables are present in their original form.

What are the main issues?

Synthetic data raises a unique set of privacy issues. The same concerns with traditional de-identification techniques remain, but with different details. However, synthetic data also raises new considerations given that it is an enabler of AI/ML systems.

- *It may be possible to re-identify individuals.* Like all de-identification techniques, synthetic data must be implemented properly, with sufficient measures to protect individuals' identities. If not, individuals can be re-identified.[9] Specifically, if the generative model learns the statistical properties of the source data too closely or too exactly, i.e., if it "overfits" the data, then the synthetic data will simply replicate the source data. Also, outliers in the source data may be vulnerable to linkage attacks if they are not removed or addressed in advance of the generation.

- *It may exacerbate biases in AI/ML systems*. The promise of synthetic data is that it will make big datasets more widely available for the purposes of training and validating AI/ML systems. However, it doesn't address the main issue with training data, namely that it may contain historical or other types of biases that would then be learned and ultimately reified in the AI/ML systems they helped to create. Indeed, by enabling and facilitating the development of AI/ML systems, synthetic data may unwittingly exacerbate the problem.

- *It may introduce its own biases*. The use of AI/ML tools to derive a generative model from the source data decreases the number of assumptions that need to be made about the underlying distributions of variables and correlations among them. However, the process is still not completely neutral. Other factors may influence it.

[9] See Theresa Stadler, Bristena Oprisenu and Carmela Troncoso, "Synthetic Data – Anonymisation Groundhog Day," https://arxiv.org/abs/2011.07018.

For example, the way the source data is pre-processed and transformed to make it accessible to the generative model may affect its properties. Also, the choice of generative model may play a role. Finally, which metrics are chosen to assess the quality of the synthetic data may emphasize certain statistical properties over others.

- *It may lead to reputational damage*. De-identification has typically considered the issue of incorrect re-identifications or "false positives" as out of scope. However, with synthetic data, given the potential broad amount of sharing it entails as well as the fact that the data look real, the issue is arguably more pressing. Also, the widespread use of social media platforms to promote "disinformation" underscores the danger of releasing "fake" data. If synthetic data happens to contain the same set of quasi-identifiers as an individual in the population, it may lead to significant reputational damage. Telling the individual that the data was fake (something they already know) may provide little comfort after the rumours have started and the online campaign is in full swing.

MESSAGES CLÉS / KEY MESSAGES:

Dr. El Emam provided nine interview questions in advance. Most are policy-based. However, one in particular is technical in nature (question 6).

The purpose of the questions is to guide the discussion, not necessarily constrain it. Nonetheless, there is some ambiguity in the questions with respect to regulatory context. For some, it is not clear whether the question is asking for the Commissioner's view on synthetic data *in general* or with respect to *current* privacy laws. I have attempted to minimize this ambiguity by focusing on the approach to de-identification in Bill C-11 and the OPC's comments in our submission.[10] I also drew on our AI consultation work, notably our proposed regulatory framework.[11] In any event, Dr. El Emam should be able to facilitate a better comprehension.

Questions

1. Have you had experience with synthetic data in your jurisdiction? For example, entities asking your office for advice on generating or using synthetic data, or synthetic data being a part of investigations?

- **No**, as of yet the OPC has not received any consultation requests on the topic of synthetic data, nor has it come up within the context of an investigation.

---

[10] OPC, "Submission of the Office of the Privacy Commissioner of Canada on Bill C-11, the *Digital Charter Implementation Act*, 2020," https://www.priv.gc.ca/en/opc-actions-and-decisions/submissions-to-consultations/sub_ethi_c11_2105/.
[11] OPC, "A Regulatory Framework for AI: Recommendations for PIPEDA Reform," (Nov 2020), https://www.priv.gc.ca/en/about-the-opc/what-we-do/consultations/completed-consultations/consultation-ai/reg-fw_202011/

2. Training a generative model is a form of data use or processing. Does this use require additional specific consent from data subjects or would the fact that it is a privacy protective technology that enhances the rights of the data subjects mitigate against that? Despite ambiguity in some statutes across the country, current practice thus far has been to treat the creation of non-identifiable information as a form of processing that does not require additional consent.

- Current federal privacy laws do not have this level of sophistication. Under PIPEDA, there are exceptions to consent that can allow personal information to be used for statistics or scholarly research, however such exceptions are very limited and are not optimized for an AI environment. But **future laws should continue with the approach taken in Bill C-11 and relax consent requirements** when personal information is de-identified in certain circumstances.
- Bill C-11 exempted organizations from obtaining an individual's knowledge or consent to de-identity their personal information.[12] However, it also placed limits on uses and disclosures of de-identified information.[13] In effect, this tied the need for consent to specific purposes.
- From a policy perspective, I agree with this approach.
- Synthetic data reduces the identifiability of information, but it also increases the potential for such information to be used or disclosed in ways which could have significant impacts on individuals' rights. Thus, in a sense it both increases and decreases privacy.
- A binary "consent / no consent" approach to synthetic data is too simplistic. A more nuanced approach is needed, where organizations have flexibility to generate and use synthetic data, but under certain conditions (see question below).

3. Should synthetic data be regulated under privacy laws, and if so, how? The implications of regulating fake data could arguably be quite impactful unless only certain types of synthetic data are regulated. How would we define these carve-outs, if any?

- **Yes, synthetic data should be regulated**.
- Synthetic data is not a "silver bullet." The same concerns with traditional de-identification techniques remain, but with different details. However, synthetic data also raises new considerations given that it enables and facilitates the development of AI/ML systems. We see four concerns in general:
    1. It may be possible to re-identify individuals.
    2. It may exacerbate biases in AI/ML systems.
    3. It may introduce its own biases.

[12] Bill C-11, s. 20: "An organization may use an individual's personal information without their knowledge or consent to de-identify the information."
[13] Bill C-11, s. 21 and s. 39.

4. It may lead to reputational damage.

- The only question is in what type of law (or laws) should regulation happen and the details of that law.
- Bill C-11 would have regulated de-identified information. It would have limited uses to certain **legitimate business practices**, such as an organization's internal research and development or prospective business transactions. It would have limited disclosures to those made for a **socially beneficial purposes** and to certain public or prescribed entities, such as health care institutions.
- By treating de-identified information as "personal information," Bill C-11 prevents this information from falling outside the scope of the law. Given the **ever-present potential for de-identified information to be re-identified**, as well as **the potential for such information to be used in ways which could have significant impacts on individuals' rights**, it is important that organizations clearly understand that, though additional flexibility is granted for certain uses, privacy legislation will remain in effect when de-identified information is used.
- I believe that Bill C-11 struck an appropriate balance with respect to de-identification, encouraging innovation by providing flexibility to organization while maintaining necessary controls and oversight.
- However, it is also interesting to note the approach the Europeans are considering.
- The EU recently put forward a draft *Artificial Intelligence Act*, with the goal of laying down "harmonised rules for the placing on the market, the putting into service and the use of [AI] systems in the Union."[14]
- If passed, the EU's proposed *Artificial Intelligence Act* would regulate synthetic data through rules and requirements with respect to training, validation and testing data sets. For example:

  > "Training, validation and testing data sets shall be **relevant, representative, free of errors and complete**. They shall have the **appropriate statistical properties**, including, where applicable, as regards the persons or groups of persons on which the high-risk AI system is intended to be used. These characteristics of the data sets may be met at the level of individual data sets or a combination thereof"[15] (emphasis added).

- At the end of the day, the question is **what are individuals' reasonable expectations** with respect to the generation, use and disclosure of synthetic data? In my view, Bill C-11 and the EU's proposed *Artificial Intelligence Act* address this question well.
- I will add that in our consultation work on AI in the winter of 2020, my Office proposed a regulatory framework that would allow the benefits of AI to be better achieved. We proposed that a new privacy law in Canada must include exceptions

[14] EU *Artificial Intelligence Act*, Art. 1 (a).
[15] Ibid. Art. 10 (3).

to consent for the use of personal information for research and statistical purposes, compatible purposes, and legitimate commercial interests purposes.

- To use information without consent for research or statistical purposes, and to the extent possible for legitimate commercial interest purposes, we noted that there should be a requirement for information to first be de-identified. And as part of this requirement, we said that **the law should prohibit re-identification** when personal information is de-identified, and the practice should be subject to financial penalties, similar to the approach Quebec's new privacy legislation. These measures are in recognition of the fact that de-identification, even properly implemented, does not negate all risk.

4. Is a data custodian able to delegate the creation of synthetic data to a third party (a sub-contractor)? What conditions would apply under these circumstances?

- **Yes, so long as appropriate contractual controls or other means are in place** to provide a comparable level of protection while the information is being processed by a third party.
- Contractual measures should, at a minimum:
    - Define the specific elements of personal information being shared;
    - Define the specific purposes for the sharing;
    - Limit secondary use and onward transfer, and;
    - Outline other measures to be prescribed by regulations, such as specific safeguards, retention periods and accountability measures.

5. Given the concerns with non-identifiable data that have been expressed in the media and by regulators recently, do you think there is a need for guidance or standards on the generation of synthetic data? How would there be assurance that these standards are being applied properly? Do we need guidance on the ethical uses of synthetic data?

- **Yes, there is a need for guidance or standards** on synthetic data, both of which could achieve a depth and degree of specificity that would be inappropriate to place within a statute.
- However, there are also a number of challenges. In particular, there are virtually no regulatory precedents specific to synthetic data. The technology has advanced rapidly, but best practices are still emerging. From a regulatory perspective, many questions are still unanswered. For example, which privacy or utility metrics should be used to evaluate the efficacy of synthetic data, in which circumstances and at what thresholds?
- To ensure guidance / standards are being followed, there are three approaches to consider: (1) complaint-driven investigations; (2) proactive audits; and (3) penetration / motivated intruder tests, where an authorized agent attempts to re-

identify or draw inferences about individuals within a synthetic dataset using other (possibly publicly available) information.

- Each has pros and cons. However, there has perhaps been too great an emphasis placed on investigations in the past. To complement compliance investigations, it may be time to rely more on audits and/or penetration / motivated intruder tests.
- Guidance on ethical uses of synthetic data would presumably be subsumed under guidance on the responsible use and development of AI.
- There is also a need for such guidance; however, best practices are more established in this field.

6. There is growing evidence that machine learning models can be attacked to recover part or all of the training data. This has resulted in some saying that machine learning models trained on personally identifying data should be treated as personal information when they are used and disclosed. However, because synthetic data is not identifiable then machine learning models trained on synthetic data would not need to be treated as personal information. What are your thoughts on this perspective?

- This is an active area of research, so **ultimately the jury is still out** on whether or to what extent synthetic data protects against privacy attacks on AI/ML models.
- There are two privacy attacks to consider:
    1. "Model inversion attacks" where an attacker infers additional personal information about an individual who is included in the training data; and
    2. "Membership inference attacks" where an attacker learns whether a given individual was present in the training data.
- Both attacks work by observing the inputs and outputs of the AI/ML model and exploiting information leakage in the confidence scores provided alongside the model's prediction.
- In general, it appears that synthetic data, by itself, would not be sufficient to protect against *all* privacy attacks on AI/ML models.
- This is because "overfitting is not the only factor that causes a model to be vulnerable to membership inference. The structure and type of the model also contribute to the problem."[16]
- However, in general, it appears that synthetic data is better at protecting against model inversion attacks than membership inference attacks.
- On the issue of whether synthetic data should be treated as personal information, I would repeat what I said earlier, that **synthetic data should be regulated** and subject to certain defined exceptions, with strong penalties for re-identification. The law should incentivize and not overly constrain responsible use of privacy

[16] Reza Shokri, Marco Stronati, Congzheng Song, Vitaly Shmatikov, "Membership Inference Attacks Against Machine Learning Models," *2017 IEEE Symposium on Security and Privacy*, p. 13.

preserving techniques that are in the interests of society. However, since re-identification remains a possibility, de-identified data (including synthetic data) must remain within the scope of a new privacy law.

7. Do data custodians need to inform data subjects if they use their data to create synthetic datasets (this would be a form of transparency rather than consent)?

- At present, if data is properly de-identified and aggregated, then **Canada's federal privacy laws likely do not apply**.
- However, a future law should create **enhanced transparency and accountability requirements** after organizations de-identify information, which is done in other jurisdictions.
- The Canadian business community has emphasized the inadequacy of the consent model and has advocated for transparency and accountability to play a larger role. However, a shift to greater reliance on accountability means greater latitude or freedom for organizations to use personal information, sometimes in dubious ways. Therefore, this approach should be accompanied by a greater role for the regulator, to ensure accountability is demonstrated and ultimately protects the rights of individuals.
- **Demonstrable accountability** must include a model of assured accountability pursuant to which the regulator has the ability to proactively inspect an organization's privacy compliance.
- In terms of transparency, Bill C-11 required organization to implement a privacy management program that included policies, practices and procedures respecting "the development of materials to explain the organization's policies and procedures put in place to fulfill its obligations under this Act."[17]
- It is also important to note the importance of **public consultation** as part of the process to establish and maintain **trust**.
- The GDPR requires organizations to consult with data subjects or their representatives as part of a Data Protection Impact Assessment, "where appropriate."[18]

8. Can synthetic data be disclosed to anyone given that it is not identifiable information?

- At present, if data is properly de-identified and aggregated, then **Canada's federal privacy laws likely do not apply**.
- However, as the Courts in Canada have ruled, "information will be about an identifiable individual where there is a serious possibility that an individual could be identified through the use of that information, alone or in combination with other

[17] Bill C-11, s. 9(1)(d).
[18] GDPR, Art. 35(9).

information."

- This suggests that pseudonomized or de-identified data could be personal information under privacy law and subject to all its provisions.
- Thus the answer of whether synthetic data as a technique of de-identification can be disclosed to anyone for any purpose is not straightforward.
- What is clear, however, is our **current federal privacy laws are outdated** and should be replaced with a more modern approach to the disclosure of de-identified data, such as in Bill C-11.

9. Can synthetic data be used for any purpose given that it is not identifiable information?

- Refer to question 8.
- However, in addition to Bill C-11, it is important to recall the EU's proposed *Artificial Intelligence Act*, given that the question is specific to the use of synthetic data.
- The EU's proposed *Artificial Intelligence Act* would require training data to be "relevant, representative, free of errors and complete" as well as having "the appropriate statistical properties."

CONSULTATIONS: PRPA, GA, Compliance

PIÈCES JOINTES / ATTACHMENTS:

Dr. El Emam's contributions grant proposal: "A Pan-Canadian Descriptive Study of Privacy Risks from Data Synthesis Practices within the Evolving Canadian Legislative Landscape" - https://officium/_layouts/15/OPC.Officium/Utilities/OfficiumIDLookup.aspx?id=7777-6-490022

Interview background materials - https://officium/_layouts/15/OPC.Officium/Utilities/OfficiumIDLookup.aspx?id=7777-6-490018

Legal analysis excerpt from Dr. El Emam's book, *Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data* - https://officium/_layouts/15/OPC.Officium/Utilities/OfficiumIDLookup.aspx?id=7777-6-490021

APPROBATION / APPROVAL:

Rédigé par / Prepared by	**Date**	**Révisions / Revisions**
David Weinkauf, Senior IT Research Analyst Technology Analysis Directorate		

Approuvé par Directeur / Approved by Director **Date**

8/2/22

Martyn Turcotte, Director
Technology Analysis Directorate

Approuvé par sous-commissaire / Approved by Deputy Commissioner **Date**

Gregory Smolynec, Deputy Commissioner
Policy and Promotion Sector

Smolynec, Gregory
c=CA, o=GC, ou=OIC-CI,
cn=Smolynec, Gregory
2022.02.08 18:44:05
-05'00'

8/2/22

Approuvé par Commissaire / Approved by Privacy Commissioner **Date**

☐ J'approuve les mesures proposées. / I agree with the proposed recommendation(s).

☐ Je ne suis pas satisfait de ces recommandations pour les raisons suivantes. / I do not agree with the proposed recommendation(s) for the following reason(s):

Commentaires ou des instructions supplémentaires / Additional comments or instructions:

Daniel Therrien
Le commissaire à la protection de la vie privée / Privacy Commissioner

DISTRIBUTION: Commissioner

A Pan-Canadian Descriptive Study of Privacy Risks from Data Synthesis Practices within the Evolving Canadian Legislative Landscape

Proposal for OPC Contributions Program 2021-2022

Khaled El Emam, PhD
5 February 2021

Abstract

Data synthesis is rapidly emerging as a practical privacy enhancing technology (PET) for sharing data for secondary purposes. However, the strengths and weaknesses of this emerging technology are not fully appreciated and need to be evaluated. As well, we need to develop an understanding of how data synthesis would be treated under various privacy regimes in Canada. In keeping with this year's Contributions Program theme of *Protecting Privacy in an Increasingly Digital World,* this project aims to provide a detailed overview of data synthesis as a PET used to facilitate data sharing within the Canadian context. It is intended to help Canadian organizations understand what data synthesis is, and also to provide an assessment of contemporary methods and technologies and how they can be applied under current and proposed regulatory regimes.

The proposed project will consist of three main research phases:

1. An overview of data synthesis (environmental scan / literature review)
2. A legal analysis of data synthesis under Part I of PIPEDA and the *Consumer Privacy Protection Act* (CPPA) component of the *Digital Charter Implementation Act, 2020* (Bill C-11)
3. Perspectives of Canadian regulators on data synthesis (interview study)

The combination of three main areas of coverage will provide a picture for the Canadian federal private and public sectors of synthetic data applications, privacy risks, perspectives on the regulation of synthetic data, and how to manage these risks. Most importantly, this research project will assess: i) whether PIPEDA and the proposed provisions of the CPPA adequately address data synthesis as a PET to protect individual privacy; ii) identify if there are gaps in both PIPEDA and in the legislative proposal, and the nature of such gaps; and iii) propose solutions to "close the gaps". These solutions may be applied to inform comment by the OPC and others on the CPPA (or any other legislative proposal to amend/repeal Part I of PIPEDA) to ensure that Canada's new privacy regime enables the use of data synthesis as an important technology to protect privacy in an increasingly digital world.

Table of Contents

1. Basic Information

Principle Investigator (applicant)	Khaled El Emam, PhD Children's Hospital of Eastern Ontario (CHEO) Research Institute Electronic Health Information Laboratory (EHIL) 401 Smyth Road, Ottawa, Ontario, Canada K1H 8L1 Tel: 613-737-7600 ext. 4147 Fax: 613-737-6504 Email: kelemam@cheo.on.ca
Project Administrator	Elizabeth Gillis Children's Hospital of Eastern Ontario (CHEO) Research Institute 401 Smyth Road, Ottawa, Ontario, Canada K1H 8L1 Tel: 613-737-4165 Fax: 613-738-4875 Email: egillis@cheo.on.ca

2. Legal Status

The Children's Hospital of Eastern Ontario (CHEO) Research Institute is a not-for-profit organization. The Electronic Health Information Laboratory, a research laboratory under the CHEO Research Institute and affiliated with the University of Ottawa, is also not-for-profit.

3. Organizational Background

The Electronic Health Information Laboratory (EHIL) was formed in 2005 under the CHEO Research Institute and headed by Dr. Khaled El Emam. Dr. El Emam is a Professor at the University of Ottawa, Faculty of Medicine, and a Senior Investigator at the Children's Hospital of Eastern Ontario Research Institute. He also held the Canada Research Chair in Electronic Health Information at the University of Ottawa from 2005-2015.

EHIL has a research program devoted to facilitating the sharing of health information for secondary purposes while protecting the privacy of patients and the identity of providers. EHIL develops technology to facilitate data sharing, including data synthesis, de-identification, federated analysis, and secure computation methods to allow surveillance and analysis without compromising privacy. The methods are suitable under different circumstances and constraints, from individual-level data release, to on-going surveillance, and to interactive remote analysis. The methods developed at EHIL are applicable, and have been applied to, personally identifiable data in health, finance, telecommunications, utilities, and retail. Therefore, the extent of impact of EHIL's research work has expanded beyond health data over the last decade and a half.

EHIL's research spans theoretical work (which consists of developing mathematical models, algorithms, and metrics for measuring and managing identification and re-identification risk), empirical work (evaluations of our models, algorithms, and metrics through simulations and controlled studies), applied work (evaluations on large real-world data sets), and knowledge translation (building software tools, instruments, and education through presentations, webinars, and workshops), as well as an active commercialization effort to disseminate innovations in privacy enhancing technologies globally and transition them into practice.

EHIL has made several critical contributions in our privacy research which represent important theoretical and practical improvements over the existing body of work, as well as very broad-based impact. Some notable examples are highlighted below:

- EHIL has developed a complete methodology for risk-based de-identification, including metrics and algorithms to enable risk measurement and the optimal transformation of large datasets to manage these risks. The development work for this methodology, which did receive seed funding from the OPC contributions program, resulted in a successful commercialization effort in the form of Privacy Analytics Inc., an Ottawa-based company that currently employs 125+ professionals developing and deploying de-identification solutions globally, almost all employed in Canada.
- EHIL's work has been incorporated into various guidelines and standards (for example, the de-identification guidelines from the Ontario Privacy Commissioner, the de-identification certification program from HITRUST, the Council of Canadian Academies report on sharing health data, and the US National Academies report on the sharing of clinical trial data). This work has also informed various court cases in the US and Canada as well legislative efforts at the municipal, provincial, state, and federal levels.
- The development of methods and guidance for the broad sharing of clinical trial data has been informed significantly by EHIL's work, as evidenced by the guidance published by the European Medicine's Agency under their Policy 0070 [1] and the Health Canada guidance on the public release of clinical information [2]. Both of these transparency initiatives have resulted in making millions of pages of clinical reports available for secondary analysis over the last 4 years.

A key principle for how EHIL operates is that it is multi-disciplinary to ensure that we are able to solve complex problems. Our collaborations include working with technologists, legal professionals, policy makers, domain experts (e.g., health and finance), and regulators.

4. Previous Financial Support

The Electronic Health Information Laboratory, under the CHEO Research Institute, received $45,000 in funding under the 2006-2007 Contributions Program for the research project titled *Pan- Canadian De-identification Guidelines for Personal Health Information*. The purpose of the project was to examine the

risks of re-identification of anonymized Personal Health Information when the data is combined with information from public databases or with inferential data (for example the predicting of gender and year of birth from first names and graduation years). The research resulted in a report with a recommended decision-making process for anonymizing a data set, with detailed considerations for different quasi-identifiers.

EHIL also received $36,000 in funding under the 2010-2011 Contributions Program for the research project titled Managing the Risk of Re-identification for Public Use Files. The purpose of the project was to write a report that discussed in detail the principles, metrics, and methods that can be used to manage the privacy risks associated with disclosing data, and to ensure that the probability of re-identifying individuals in publicly disclosed files is low and that the probability of discovering sensitive information about them is low.

That specific report provided methods for determining what "low" should be and what would be appropriate levels of access restrictions on public data. The objective was to give data custodians the tools to make decisions about the best way to disclose this data, but also ensure that the privacy of individuals is protected. The report was combined with other research from EHIL into the book *Guide to the De- Identification of Personal Health Information*, published by CRC Press in 2013 [3].

Most recently, EHIL received $50,000 in 2016 for the development of two online learning courses on Data Privacy and Data Anonymization intended for the private sector. The first course on Data Privacy was an introductory course intended to educate data recipients on the legal framework and disclosure risks for personally identifiable information. The second course on Data Anonymization was a more in-depth course intended for practitioners looking to apply the best guidelines and strategies in the de-identification of personally identifiable information. The courses were offered free of charge online via the EHIL website, and we received positive feedback from stakeholders on the relevance and usefulness of the material in training both internal staff and external recipients of data.

5. Project Team and Resources

5.1. Khaled El Emam

Dr. Khaled El Emam is a Senior Scientist at the Children's Hospital of Eastern Ontario (CHEO) Research Institute and Director of the multi-disciplinary Electronic Health Information Laboratory (EHIL) team, conducting academic research on data synthesis, de-identification and re-identification risk measurement, secure computation, and federated analysis. He is also a Professor in the School of Epidemiology and Public Health at the University of Ottawa.

He currently sits on a number of committees and advisory boards, including the Technical Anonymization Group of the European Medicines Agency, and the privacy advisory group of the UN Global Pulse, as well as other groups developing guidelines and training for sharing health data.

Khaled (co-)founded six companies focused on data management, data analytics and privacy preserving technologies, and also invests in companies developing digital health technologies. He currently sits on the board of Replica Analytics and on the board of Canary Medical. He has worked in technical and management positions in academic and business settings in Canada (CRIM, Montreal; McGill University, Montreal; NRC, Ottawa; Trialstat, Ottawa; Privacy Analytics / IQVIA, Ottawa; uOttawa, Ottawa; CHEO RI, Ottawa), Germany (Fraunhofer Institute, Kaiserslautern), England (Toshiba International, London),

Scotland (Honeywell, Glasgow) and Japan (Yokogawa Electric, Tokyo).

In 2003 and 2004, Khaled was ranked as the top systems and software engineering scholar worldwide by the Journal of Systems and Software based on his research on measurement and quality evaluation and improvement. He previously held the Canada Research Chair in Electronic Health Information at the University of Ottawa. He has a PhD from the Department of Electrical and Electronics Engineering, King's College, at the University of London, England.

5.2. Anita Fineberg

Anita Fineberg is a sole practitioner and consultant in Toronto specializing in the areas of privacy, access to information, data security and information management, predominantly in the health sector. She provides advice to both public and private sector entities on the privacy and security requirements of EHR/EMR systems, mobile applications for health, remote condition monitoring and data governance. Anita is a passionate advocate for the principles of *Privacy by Design*, believing that the adoption of a proactive approach to privacy best serves her clients and their patients.

She has practiced privacy law for over 30 years, having spent seven years with the Office of the Information and Privacy Commissioner/Ontario and three years as counsel to the Ontario Ministry of Health and Long-Term Care. Prior to establishing her own practice, Anita was Corporate Counsel and Chief Privacy Officer to IMS Health Canada, the Canadian affiliate of IMS Health (now IQVIA). She holds a B.A. (Hons. Psychobiology) degree from Queen's University, a LL.B. degree from the University of Toronto, and is a designated Certified Information Privacy Professional/Canada.

Anita is a member of the Privacy and Advocacy Sections of the Canadian Bar Association and the Privacy, Health, Technology and Administrative Law Sections of the Ontario Bar Association. She is also a member of the International Association of Privacy Professionals, the Canadian Association of Management Consultants, as well as HIMSS (the American Healthcare Information and Management Systems Society) and COACH (Canada's Health Informatics Association). Anita completed a three-year term as the public member on the Health Technology Expert Review Panel (HTERP), an advisory body to the Canadian Agency for Drugs and Technologies in Health (CADTH).

Anita is a frequent speaker, workshop leader and author on the privacy and security of electronic health records systems and technology and is an adjunct professor at the Ryerson University Chang School of Continuing Education for which she developed and teaches courses in Health Information Privacy and Access as well as Applied Concepts in Privacy and Access. She is also the Program Advisor for the Chang Certificate in Privacy, Access and Information Management. Anita is a founding member of the HealthLawyerNetwork (HLN) http://www.healthlawyernetwork.ca/.

5.3. Elizabeth Jonker

Elizabeth Jonker is Research Coordinator and Privacy Officer of the Electronic Health Information Laboratory at the CHEO Research Institute. Elizabeth has over thirteen years of experience in privacy research, contributing to numerous research projects and co-authoring fifteen articles published in academic journals. She is a member of the IAPP and has been a Certified Information Privacy Professional (CIPP/C) since 2012.

Elizabeth has prior experience in program coordination and facilitation with the City of Ottawa. She holds an Honours BA from the University of Ottawa from which she graduated magna cum laude.

5.4. Potential Conflicts of Interest and Their Management

Dr. Khaled El Emam is a co-founder and Director at CANON (the Canadian Anonymization Network), co-founder of Replica Analytics (which provides synthetic data services), sits on the board of a number of companies including Replica Analytics and Canary Medical (a medical device company), and actively invests in digital health technology companies. He is also an advisor (on data protection technologies and AI) to a number of companies and government agencies in Canada and globally.

There is a strong mandate and expectation from Canadian research funding agencies that funded researchers should work with industry and to commercialize the results of research work, as appropriate. Potential conflicts are managed by ensuring that research projects go through a research ethics review board and all funding proposals to external funding bodies are reviewed by an internal institutional committee. There is also a mandatory annual disclosure of Conflicts of Interest within the CHEO Research Institute which results in additional oversight if there are concerns by institutional leadership. Failure to declare COIs can result in termination of affiliation with the research institute, and therefore the Conflict of Interest issue is taken seriously.

6. Project Overview

6.1. Project Summary

Data synthesis is rapidly emerging as a practical privacy enhancing technology (PET) for sharing data for secondary purposes. The rate of adoption of data synthesis has been growing steadily with a marked increase in 2020. Applications of data synthesis technologies vary from AI and machine learning model building and analyses with the synthetic data, open data and open science initiatives, to technology evaluation and software testing applications.

When the coronavirus pandemic took hold of the world early in 2020, the necessity to share patient data broadly but also safely in order to fight the spread of the virus became paramount. As a result, interest in data synthesis has increased as a means to facilitate sharing of health data while mitigating the risks to patient privacy.

However, the strengths and weaknesses of this emerging and increasingly adopted technology are not fully appreciated and need to be evaluated, and we need to develop an understanding of how data synthesis would be treated under various privacy regimes in Canada. This project aims to provide a detailed analysis of data synthesis in a Canadian context. It is intended to serve an educational purpose to help readers understand what data synthesis is, how it is applied in practice, and also to provide an assessment of contemporary methods and technologies. It is also designed to assess how this PET can be treated under current regulatory regimes, identify gaps and propose how it could be treated under proposed regulatory regimes for maximum privacy protection.

The proposed project has three main research phases. Note that the projected output is intended to be accessible to a broad (non-technical) audience.

An overview of data synthesis (environmental scan / literature review)

This phase of the project will document the history of data synthesis and provide an overview of current methods (statistical, machine learning, and deep learning) for data synthesis and their applications. Key use cases in multiple verticals will be used to illustrate the potential value of synthetic data and the kinds of problems it can potentially solve. This section will also provide an evidence-based critical appraisal of the strengths and weaknesses of data synthesis, as far as they are documented in the literature, from the perspectives of data utility and the risks of identity disclosure, and compare the findings to standard/current de-identification methods. In particular, we will examine and critique the models that have been used for assessing privacy risks and the results of these assessments thus far.

A legal analysis of data synthesis under PIPEDA and the CPPA

Data synthesis is an approach for rendering personal information to be non-personal. In a previous analysis we examined three limited questions in the context of the GDPR, CCPA (California), and HIPAA: (a) is the use of the original (real) data set to generate and/or evaluate a synthetic data set restricted or regulated under the law, (b) is the sharing of the original data with a third- party service provider to generate synthetic data restricted or regulated under the law, and (c) does the law regulate or otherwise affect (if at all) the resulting synthetic data set?

For this project, we plan to conduct a much broader, four-part legal analysis. We will: i) initially identify the privacy risks that exists with respect to the use of synthetic data as a PET; ii) assess these risks against

PIPEDA and the proposed provisions of the CPPA; iii) conduct a "gap analysis" based on the risk assessment; and iv) provide recommendations on how these gaps may be addressed to ensure that PIPEDA reform, whether as currently proposed in the CPPA or otherwise, will ensure the appropriate regulation of data synthesis and enable its many positive uses while upholding individuals' privacy rights.

Perspectives of Canadian regulators on data synthesis

We plan to contact federal, provincial and territorial privacy commissioners and interview staff within their offices to get their perspectives on the risks and benefits of using data synthesis to facilitate access to data for secondary purposes. The interview script will be developed with a panel of privacy experts. The findings will provide a pan-Canadian perspective on how regulators view this emerging approach.

The combination of three main areas of coverage as outlined above will provide a somewhat comprehensive picture for the public and private sectors in Canada on the application of synthetic data, its risks and benefits, and perspectives on its regulation.

6.2. Relevance to OPC Priorities

This year's Contributions Program theme is *Protecting Privacy in an Increasingly Digital World*. The current relevancy of the theme is anchored in the pandemic and Canadians' increased online interactions:

> *Examples of this phenomenon speak for themselves. Telemedicine provided during a pandemic has undeniable advantages, but when it is offered via online platforms by private-sector organizations, the risk to the confidentiality of health information increases. Distance education creates similar risks. While the challenges of our times are extensive, we cannot simply set rights aside, and this includes the right to privacy.* ***With the theme of protecting privacy in an increasingly digital world, we are asking funding applicants to reflect on how laws and best practices must be applied, in the unique and historical context of the pandemic.***

Data synthesis as a PET is not only a key mechanism to manage the risks of such digital interactions, but also integral to the protection of privacy in an increasingly digital world, which, as pointed out by the Privacy Commissioner is at the core of the fourth industrial revolution.

Our proposal is specifically focused on describing both the state of the art and of practice of synthetic data with respect to privacy, and the effectiveness of data synthesis to minimize the risks to the protection of personal information. The legal analysis of the treatment of synthetic data in PIPEDA and the proposed CPPA would inform the development of a regulatory framework.

7. Project Description

7.1. Interest in Data Synthesis Has Been Growing Rapidly

Interest in synthetic data has been growing quite rapidly over the last few years. This has been driven by three simultaneous trends. The first is the demand for large amounts of data to train and build artificial intelligence and machine learning (AIML) models. The second is recent work that has demonstrated effective methods to generate high quality synthetic data. Third, trust in contemporary de-identification methods has been eroded by repeated claims of successful re-identification attacks on anonymized data [4]–[10], reducing public and regulator confidence in this approach [10]–[18]. There is a need for more advanced methods to create non-identifiable data.

These trends have resulted in the growing recognition that synthetic data can solve some difficult problems quite effectively, especially within the AIML community. Groups and businesses within companies like NVIDIA, IBM, and Alphabet, as well as agencies such as the US Census Bureau, have adopted different types of data synthesis to support model building, application development, and data dissemination.

When the coronavirus pandemic took hold of the world early in 2020, the necessity to share patient data broadly but also safely in order to fight the spread of the virus became paramount. CBC news coined this "humanity's first data-driven pandemic" [19], citing an unprecedented level of global data sharing concerning the novel coronavirus and its spread. And there are numerous websites dedicated to tracking the spread of the disease [2], [3], [4], [5], [6] as well as mutations of the virus [19], [25].

Many argue that data is exactly what is needed to fight the pandemic and are pushing for more open sharing of data with researchers, health care providers, and public health organizations [26]–[28]. This will allow us to better understand the disease and work more quickly toward treatments and vaccines. Also, AI methods have been effectively applied to analyze COVID-19 data and provide projections, but these methods require large volumes of data [29].

With broad sharing of information, heightened by the data requirements of AIML methods, come risks to privacy that must also be managed. Patients who have tested positive for the virus may run the risk of being ostracized and discriminated against. For example, in South Korea where a great deal of information was released publicly about patients who tested positive for the virus, including where those individuals have been and when, people are "managing to connect the dots and identify people"[35]. Results of a survey conducted by a team at Seoul National University's Graduate School of Public Health found that participants were more afraid of "Criticisms and further damage they may suffer from being infected" [35] than of contracting the virus.

As a result, interest in PETs such as data synthesis has grown as a means to facilitate broader sharing of health data while helping to address the privacy risks. In fact, a number of recent efforts have made large COVID-19 datasets available through data synthesis. The Clinical Practice Research Datalink (CPRD) database in the UK has made available a COVID-19 symptoms and risk factors synthetic dataset based on primary care encounters in the UK [30], [31] In the US, the NIH's N3C is also developing synthetic datasets for broader sharing with researchers [32], [33]. In South Korea, the Health Insurance Review and Assessment (HIRA) service (the national health insurer) provides synthesized data to researchers. Researchers may then submit the analysis code they developed and tested on the synthetic data to the agency to be run on the real data. As a result, the researchers are able to obtain results from both the real data as well the synthetic data [34]. There are a number of other efforts in progress to share synthetic COVID-19 data; interest in this privacy protective approach for sharing data is growing.

Further examples of the growing interest in data synthesis are provided below. See the sidebar demonstrating recent trends in research and startups (many have been formed in the last 18-24 months).

Growth in Data Synthesis Literature

To explore the increase in interest in data synthesis, we conducted a preliminary literature search in the Scopus database to find out how many publications relating to data synthesis techniques and applications were published in 2020. Scopus is an extensive database that indexes content from 24,600 active titles and 5,000 publishers, covering the subject areas of computer science, engineering, medicine (including medical informatics), and the natural sciences. Our search strategy is outlined in more detail in the Appendix to this proposal.

Our search returned a total of 3025 publications related to synthetic data that were published in 2020. The subject areas of these publications ranged from computer science and engineering to medicine to biochemistry, genetics and molecular biology, with many areas in between. For popular methods currently used to generate synthetic data, the search returned 452 results related to Generative Adversarial Networks, 86 related to Bayesian Networks, and 36 related to Copulas.

Therefore, the level of research and technology development in the area of synthetic data is quite large and continuously growing.

More Data Synthesis Startups

There have also been significant investments in startups focused on developing data synthesis technology. Some examples of recent startups that have been funded are below. That there is investment flowing into these companies and in transitioning this technology into practice suggests that there is a growing market (the investors must have determined that there will be substantial returns).

Company Name	Location
Replica Analytics	Canada
edgecase.ai	USA
mostly.ai	Austria
MDClone	Israel
Realsynth	Germany
AI.Reverie	USA
syntho.ai	Netherlands
synthesized.io	UK
WeData	France
Syntegra	USA
GenRocket	USA
Hazy	UK
statice.ai	Germany
Facteus	USA
tonic.ai	USA
Virtusa	USA
Datomize	Israel

The US Census Bureau has, at the time of writing, decided to leverage synthetic data for some of the most heavily used public datasets, the 2020 decennial census data. For their tabular data disseminations, they will create a synthetic dataset from the collected individual-level census data and then produce the public tabulations from that synthetic dataset. A mixture of formal and non-formal methods will be used in the

synthesis process [36].

This, arguably, further demonstrates the continued large scale adoption of data synthesis for one of the most critical and heavily used datasets available today. We may be past the point of asking whether synthetic data will become more widely adopted, and should be more focused on what frameworks, policies, guidelines, and regulations are required to ensure the responsible use of synthetic data.

7.2. Use Cases for Synthetic Data

Historically, access to data for AIML projects has been problematic in practice. The General Accountability Office [37] and the McKinsey Global Institute [38] both note that accessing data for building and testing AIML models is a challenge for their adoption more broadly. A Deloitte analysis concluded that data access issues are ranked in the top three challenges faced by companies when implementing AI [39]. A recent survey by O'Reilly highlighted the privacy concerns of companies adopting machine learning models, with more than half of companies experienced with AIML checking for privacy issues [40]. At the same time, the public is getting uneasy about how their data is used and shared, and regulatory scrutiny of secondary uses and disclosures of data is growing.

To address such concerns, the Privacy Commissioner has made recommendations for the appropriate regulation of AI in PIPEDA [41], only some of which have been addressed in the CPPA.

Data synthesis is gaining momentum as a potential solution to enable privacy protective and responsible access to large amounts of data for the AIML community.

At a conceptual level, synthetic data is not real data but is data that has been generated from real data and that has the same statistical properties as the real data. This means that if an analyst works with a synthetic dataset they should get similar analysis results to what they would get working with real data. The degree to which a synthetic dataset is an accurate proxy for real data is a measure of utility.

Data in this context can mean different things. For example, data can be structured data as one would see in a relational database. Data can also be unstructured text, such as doctors' notes, transcripts of conversations among people or with digital assistants, or online interactions by email or chat. Furthermore, images, videos, audio, and virtual environments are additional types of data that can be synthesized.

There are two types of synthetic data depending on whether they are generated from actual datasets or not.

The first type is synthesized from real datasets. This means that the analyst will have some real datasets and then builds a model to capture the distributions and structure of that real data. Here structure means the multivariate relationships and interactions in the data. Then the synthetic data is sampled or generated from that model. If the model is a good representation of the real data then the synthetic data will have similar statistical properties as the real data.

For example, a data science group specializing in understanding customer behaviors would need large amounts of data to build their models. But the process for getting access to that customer data is slow and does not provide them with good enough data when it does arrive because of extensive masking, generalization, and redaction of information. Instead, a synthetic version of the production datasets can be provided to the analysts to build their models with. The synthesized data will potentially have better utility, fewer constraints put on its use, and would allow them to progress more rapidly.

The second type of synthetic data is not generated from real data. It is created by using existing models or by using background knowledge of the analyst. These existing models can be statistical models of a process, for example, developed through surveys or other data collection mechanisms, or they can be simulations. Simulations can be, for instance, gaming engines that create simulated (and synthetic) images of scenes or objects, or simulation engines that generate shopper data with particular characteristics (say, age and gender) who walk past a planned store at different times of the day.

Background knowledge can be, for example, a model of how a financial market behaves based on textbook descriptions or based on the behaviors of stock prices under various historical conditions, or it can be knowledge of the statistical distribution of human traffic in a store based on years of experience. In such a case it is relatively straight forward to create a model and sample from it to generate synthetic data. If the analyst's knowledge of the process is accurate, then the synthetic data will behave in a manner that is consistent with real world data. Of course, this only works when the phenomenon of interest is truly well understood.

As a final example, where a process is new or not well understood by the analyst and there is no real historical data to use, then an analyst can make some simple assumptions about the distributions and correlations among the variables involved in the process. For example, the analyst can make a simplifying assumption that the variables have normal distributions and "medium" correlations among them, and create data that way. This type of data will likely not have the same properties as real data but can still be useful for some purposes, such as debugging an R data analysis program or for some types of performance testing of software applications.

Our primary focus in this proposal is on synthetic data generated from real data, where the real data pertains to individuals and covers potentially sensitive information about the data subjects, for example, health data and financial transactions data. Furthermore, we will limit our investigations to structured data (as opposed to text, video, or images).

The current proposal is one step in the direction of ensuring the responsible use of synthetic data in Canada by understanding the privacy risks associated with this kind of data, how these privacy risks have been addressed to date, and proposing a regulatory framework to close the current gaps. This is particularly important given the role that non-identifiable information plays in the CPPA, for example.

The project consists of three research phases, followed by a consolidation step where the findings are converted into a summary of what is known about the privacy risks from synthetic data and some recommendations on how these risks can be managed.

7.3. Phase 1: Environmental Scan

While model-based methods for data synthesis were introduced in the early 90's [42], [43], they were based on methods borrowed from imputation (estimating missing values in data). Since then, there have been significant advances in synthesis methods, with more promising ones not requiring the specification of a model a priori, such as decision tree based approaches [44]. More recently deep learning methods have been used for data synthesis, such as Variational Auto Encoders [45], [46] and Generative Adversarial Networks [47]. The rapid advances in methodology mean that the utility of synthetic data has been improving quite rapidly, increasing the number of use cases where data synthesis can be applied.

In general, the academic literature considers that fully synthetic data does not have an identity disclosure risk, for example, because there is no unique mapping between the records in the synthetic data with the

records in the original data [48]. Reiter says that "identification of units and their sensitive data from synthetic samples is nearly impossible." [49]. Taub et al. noted that "it is widely understood that thinking of risk within synthetic data in terms of re-identification, which is how many other SDC [Statistical Disclosure Control] methods approach disclosure risk, is not meaningful" [50]. Similar views have been expressed by other authors [51]–[54], and there have been some initial empirical assessments of identity disclosure risks [55]–[58]. Furthermore, interestingly there are recent examples of research studies using synthetic data not requiring ethics review because they are considered to contain no patient information [59].

However, these broad conclusions are not necessarily accurate since data synthesis methods, especially those contemporary ones utilizing machine learning and deep learning modeling techniques, are at risk of overfitting. When the synthesis models are overfit then they will replicate the original datasets, creating a significant privacy risk. Therefore, it is still important to assess the disclosure risks from synthetic data and empirically evaluate what the likelihood is for a 1 to 1 mapping between the synthetic data and real individuals in the population. There are also other privacy models that may be relevant beyond a 1 to 1 mapping between synthetic records and real individuals (i.e., identity disclosure), such as attribute disclosure, inferential disclosure, and membership disclosure.

There are other privacy and ethical issues surrounding the use of AI and ML algorithms more generally, such as the potential for information leakage, algorithmic transparency, bias and the potential for discrimination, and other implementation related risks. These broader issues will not be addressed in this project as they are outside of the scope of our investigation. Our focus will be on the potential risks related specifically to the creation and use of synthetic data, rather than on AI and ML methods more generally.

The purpose of this environmental scan is to document the different methodologies that have been developed to evaluate the privacy risks in synthetic data, the different types of disclosure risks that have been measured and managed by scholars in this area, and their actual risk assessment results. We will also look at studies comparing synthetic data with de-identified data in terms of privacy risks as well as their relative merits. This will ground the discourse on data synthesis with real privacy assessment results.

7.3.1. Search Strategy

Articles in the computer science literature and medical informatics literature will be searched using the general terms "synthetic data" or "data synthesis", "simulated data" and "disclosure control" or "privacy risk" or "disclosure risk" or "risk analysis". Broad search terms will be used to ensure that we do not miss any relevant publications. The searches will be performed on PubMed, IEEE Xplore (the on-line library of the Institute of Electrical and Electronics Engineers) and the ACM Digital Library (the on-line library of the Association for Computing Machinery), and the records for all relevant English language articles will be obtained for further consideration. The IEEE and ACM publish and index a significant amount of the computer science and medical informatics research work. The following pre-publication repositories will also be searched: arxiv, bioarxiv, and SSRN. The resulting set of articles may be augmented with articles known to the authors, identified through targeted searches on Google Scholar (e.g., for specific authors), and articles identified through the reference lists of the included studies. Relevant technical reports and presentations may also be included. The initial search described in the appendix can be informative for refining the search strategy.

If the number of articles that meet the inclusion criteria is very large, we will use the Insightscope tool

(developed at the CHEO Research Institute) to use crowdsourcing to screen the identified articles [60]. This will allow us to efficiently narrow down large volumes at the input to the review funnel into a subset that can be analyzed in depth.

7.3.2. Analysis and Reporting

The article titles, keywords and abstracts will be screened to determine if they relate to the five questions below:

1. What are the methods that are being used to synthesize data?
2. How is data synthesis being applied in practice/what are the use cases?
3. What is the utility of synthetic data and how has that been assessed?
4. What are the privacy models that have been used to evaluate the privacy risks in synthetic data, and what are the results of empirical studies evaluating the privacy risks according to these models?
5. How is data synthesis related to and different from traditional de-identification methods?

Two independent reviewers will be involved in screening at this level to ensure accuracy.

The full text of those articles that pass the initial screening will be obtained for further review and data extraction. A template data extraction instrument based on the JBI Reviewer's Manual will be used for data extraction.[1] We will structure our review based on Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.[2]

In addition, we will support the legal review in Phase 2 by preparing summaries of information from the Environmental Scan that address the questions of "What are the privacy models that have been used to evaluate the privacy risks in synthetic data, and what are the results of empirical studies evaluating the privacy risks according to these models?" These summaries will be provided to our legal expert to inform the legal review, as well as being used to inform the final report. This information will be extracted during the review of the full text articles.

7.4. Phase 2: Legal Analysis of Data Synthesis under PIPEDA and the CPPA

This phase of the project will assess how data synthesis is currently treated under PIPEDA and its proposed treatment under the CPPA. The fact that synthetic data is not "real data" because it is not related to real people, does not mean that privacy laws are not relevant to a discussion of data synthesis. In fact, the opposite is true because data synthesis is an approach for rendering personal information to be non-personal.

The legal analysis will be conducted in a manner analogous to the conduct of a Privacy Impact Assessment (PIA) consisting of four parts: Part I will initially identify the privacy risks that exist with respect to the use of synthetic data as a PET; Part II will assess these risks against PIPEDA and the proposed provisions of the CPPA; in Part III a "gap analysis" will be conducted based on the risk assessment conducted in Part II; and

[1] See <https://wiki.joannabriggs.org/display/MANUAL/JBI+Reviewer%27s+Manual>
[2] See <http://www.prisma-statement.org/>

Part IV will provide recommendations on how these gaps may be addressed to ensure that PIPEDA reform, whether as currently proposed in the CPPA or otherwise, will ensure the appropriate regulation of data synthesis to enable its many positive uses while upholding individuals' privacy rights.

Furthermore, it is important that the definitions of non-identifiable information and de-identification in CPPA or otherwise are not based only on historical methods but will work with new methods and technologies. These definitions, and any regulations that are based on them, should be robust to the evolving landscape of privacy enhancing technologies. Our proposed analysis will inform achieving that goal.

I. Identification of the privacy risks that exist in the "lifecycle" of synthetic data.

These risks will include matters such as:

- The use of the original (real) data set to generate and/or evaluate a synthetic data set ("created the synthetic data set")
- The legal characterization of the synthetic data set
- The sharing of the original data set by the organization that created the synthetic data set with a third-party service provider to generate the synthetic data set
- The use of the synthetic data by the organization that created the synthetic data set
- The disclosure of the synthetic data to different categories of recipients (e.g., those subject to other privacy legislation, the "world at large")
- The impact of the creation, use and disclosure of synthetic data on individual privacy rights: consent, access, correction.
- The obligations of organizations to be transparent with respect to their processing of personal information

Other privacy risks will be identified related to the use of this technology, based on the privacy models that have been used to evaluate the privacy risks in synthetic data (based on the articles reviewed in the Environmental Scan conducted in Phase 1 of the Project).

II. Treatment of these risks under PIPEDA and the Proposed CPPA

This part of the legal analysis will assess the privacy risks identified in Part I as against their current regulation in PIPEDA and as proposed in the CPPA.

For comparative purposes, this analysis will consider, the three questions we previously considered in the context of the GDPR, CCPA (California), and HIPAA: (a) is the use of the original (real) data set to generate and/or evaluate a synthetic data set restricted or regulated under the law, (b) is the sharing of the original data with a third party service provider to generate synthetic data restricted or regulated under the law, and (c) does the law regulate or otherwise affect (if at all) the resulting synthetic data set?

III. Gap analysis

In this part of the legal analysis, we will present the assessment conducted in Part II in the form of a tabular gap analysis that identifies the data synthesis privacy risks, their treatment under each of PIPEDA and the proposed CPPA and the resultant legislative gaps. The gaps will be described in three categories related to the privacy risks: i) not addressed at all; ii) partially addressed; and iii) addressed but could be

improved to better protect individual privacy rights.

As is the case with Part II, our previous findings in the context of the GDPR, CCPA (California), and HIPAA will also be included.

IV. Recommendations

The legal analysis will conclude with proposed solutions on how the regulatory gaps "may be closed" in order that Canada develop a privacy regime that, in an increasing digital world, both adequately protects privacy while at the same time fostering the use of PETs such as data synthesis to foster trust and prepare the country for the fourth industrial revolution. The solutions will be drafted as an appropriate law for data synthesis that recognizes its exponentially increasing use and widespread adoption – to provide input into the government's continued consultations on Bill C-11 based on an informed understanding of the legal distinctions between data synthesis and other PETs such as de-identification.

This portion of the project will be carried out by legal expert Anita Fineberg. Anita has practiced privacy law for over 30 years and was involved in our previously funded OPC contributions project "E-Learning Course on Anonymizing data". More about Anita's experience and qualifications can be found in Section 5: Project Team and Resources.

7.5. Phase 3: Interviews with Canadian Regulators

The objective of this third phase of the project is to understand the perspectives of Canadian regulators on the same four areas outlined in the legal review: i) Identifying privacy risks associated with the use of synthetic data as a PET; ii) Assessing the risks against the current legal framework (PIPEDA and the proposed provisions of the CPPA); iii) Identifying areas where gaps are perceived in the current legal framework; and (iv) Recommendations on how these gaps may be addressed. We also seek to gather information on regulators' experiences with synthetic data within their jurisdictions (such as case studies, complaints, investigations, and queries that have come through their offices).

7.5.1. Study Design

For the interview portion of the project, we plan to use a triangulation design for this mixed methods study [61], [62] to explore participants' perspectives on the risks and benefits of using data synthesis to facilitate access to data for secondary purposes. An experienced interviewer will conduct a series of fifteen to twenty interviews with staff from the offices of federal, provincial and territorial privacy commissioners within Canada.

In terms of deciding on an adequate number of participants, there is no standard that is generally accepted. However, there are a few considerations that one should take into account when deciding on the number of participants required [63], [64]. Firstly, there is the consideration that there are a limited number of viewpoints on a topic. Having a greater number of participants does not therefore necessarily entail a greater understanding of the topic [63]. The researcher will have a sense of participants' potential viewpoints before commencing the interviews as a result of the literature review completed on the topic in phase one of the project [63]. Although opinions can vary, this would give the researcher an idea of when he/she has reached saturation; the point at which there is nothing new to be uncovered on the topic [64]. Once saturation has been reached, further interviews will no longer be necessary nor particularly useful [63], [64]. Secondly, the researcher's own ability to recall, process and understand the interviews needs to be considered [63]. As Gaskell points out, "the interviewer must be able to bring to

mind the emotional tone of the respondent and to recall why they asked a particular question" [63]. Therefore, there is a limit to the number of encounters that the researcher will be able to recall in detail. For individual interviews, Gaskell suggests that this limit would be between 15 – 25 interviews [63]. Consistent with the recommendations in the literature and with precedents, we plan to interview 15 – 20 participants, and expect to reach a point of saturation in terms of new concepts identified within that number.

The interviews will be conducted by phone / video-conference and follow an interview guide developed by the researchers in collaboration with relevant stakeholders. The duration of each interview is expected to be between thirty and forty five minutes.

7.5.2. Study Sample

For the purposes of this study, purposive sampling will be used [65]. Participants will be recruited from the offices of provincial and territorial privacy commissioners within Canada. The recruitment target is 15 – 20 individuals. Assuming a 33% drop-out rate, we aim to recruit 25 – 30 potential participants. Invitations will be sent via email and participants will self-select to participate in the study.

Individuals will be identified through the personal network of the principal investigator. We will contact the privacy (and information) commissioners directly and ask them to nominate individuals in their offices who can answer the questions. We are looking for 2 or 3 individuals from each office at most.

During our recruitment, the regulators will be complemented with members of the privacy and access law section from the Canadian Bar Association. We will contact the executive (namely, Alexis Kerr and Timothy Banks) for their help in identifying members who would be good candidates for the interviews.

7.5.3. Data Analysis

The interviews will be audio recorded and transcribed verbatim. We plan to use an approach informed by grounded theory [66], [67] to analyze the interview data. The objective of the analysis will be to understand participants' perspectives on the risks and benefits of using data synthesis to facilitate access to data for secondary purposes. Data transcription and analysis will be carried out in parallel to the interviews and will continue until a set of stable themes develops. Following each interview, the interviewer will conduct a cursory analysis of the audio recordings and field notes to determine if changes to the interview guide are required in preparation for the next interview. Question may be added, removed or reworded as required. Using the constant comparison method [66], as well as the above-mentioned constructs from the literature, we will develop a coding scheme that will embrace the themes presented in the data. Two research team members trained in qualitative research methods will independently code the transcripts using NVivo software and an inductive process. Once the members complete their independent coding, they will compare their coding and discuss it to come to an agreement. Throughout the process, codes may be modified, merged, or eliminated as required to increase the accuracy of the analysis.

7.6. Report Development

The final report will consolidate the findings from the three phases of research described above. The consolidation will focus on the same four areas as the legal review and interviews:

- Identifying the privacy risks that exists with respect to the use of synthetic data as a PET.

- Assessment of the identified risks against PIPEDA and the proposed provisions of the CPPA.
- Providing a "gap analysis" based on the risk assessment above.
- Provide recommendations on how these gaps may be addressed to inform regulatory reform.

We will present a summary of the findings based on evidence from the literature, the legal analysis, and the responses of regulators. Common themes will be identified, and special cases or situations will be highlighted. Specifically, lessons learned from current practices will be emphasized as well as identified gaps.

We anticipate a forward-looking analysis which will identify, based on the data collected during the research phases, how privacy risks are expected to evolve over time as data synthesis methods improve. Suggested or anticipated remedies to manage these risks will be discussed.

Recommendations will be developed based on the lessons learned, and a framework for managing privacy risks from the use and disclosure of synthetic data will be proposed. While the exact nature of this framework will emerge from the results of the project, the framework is expected to help make good utility data available for secondary purposes, and explain how to meet regulatory requirements and protect data subjects against disclosure risks using best known practices. It is also expected to include recommendations to address individual rights with respect to the creation and use of synthetic data, such as transparency and accountability.

The report will reflect best available knowledge today around managing privacy risks from the use and disclosure of synthetic data, and provide Canadian-specific recommendations for the application of this emerging PET.

8. Community Involvement

Four different communities will be actively engaged in this project to define its direction, and also to provide feedback on the final report. These communities reflect the private sector in Canada, the public sector in Canada, regulators, and civil society.

The principal investigator is a co-founder and director of the Canadian Anonymization Network (CANON) [3]. CANON is a not-for-profit corporation whose members include large Canadian data custodians from across the public, private and health sectors. CANON's primary purpose is to promote de-identification in Canada as privacy-respectful means of supporting innovation and leveraging data for socially and economically beneficial purposes. Because of this focus, CANON members from the private and public sectors are made up of organizations across Canada who are heavy users of data for secondary purposes.

We will leverage this relationship with CANON to involve its organizational members in refining the key questions for the various phases of the three research projects that will be performed in this project in the form of an advisory panel. This will ensure that there is a thorough review of the planned direction of the research and that it answers practical questions of relevance to the broader community within Canada.

[3] See <https://deidentify.ca/>

Private Sector AccessPrivacy; Bell; BMO; Cryptonumerics; Georgian Partners; IBM; IMS/IQVIA; Integrate.ai; Magna International; Microsoft; Moneris; National Bank; Privacy Analytics; PwC; RBC; Roche; Rogers; SunLife; Symcor, Inc.; TD Bank; Telus; TransUnion
Public and Health Sectors Alberta Health Services; Canada Health Infoway; Canadian Institute of Health Information;; Employment and Social Development Canada; Health Canada; Health Data Coalition of British Columbia; HITRUST; Metrolinx; Statistics Canada; Vancouver Coastal Health; Waterfront Toronto

Table 1: Current organizational members of CANON.

By definition, regulators across the country will be involved in this project as they will be interviewed / are part of the data collection effort.

While civil society is not expressly represented in CANON membership, we will invite specific organizations that would have an interest in this topic to participate and to provide input throughout the process, and to keep them informed. The two specific organizations that we will approach are CIPPIC and CCLA.[4]

9. Dissemination of Results / Knowledge Translation

The results of this work will be disseminated in three ways:

- We will organize a webinar to coincide with the public release of the final report and this will be recorded and made available on the lab's Youtube channel. The lab's mailing list exceeds one thousand verified individuals and its webinars typically attract 100 to 150 participants. Other social media efforts will be used to disseminate the results directly by EHIL and through partners (e.g., CANON). These include LinkedIn and Twitter posts, and blog posts related to the project.
- Presentations of this work will be made at various conferences. Specifically, EHIL has had a regular presence at the IAPP events and the Reboot events across the country. The principal investigator gives at least 30 presentations every year in various open and closed venues and the results of this work will be incorporated into that series of presentations moving forward, subject to any limitations imposed by the pandemic
- Academic articles will be written describing the work that was done. While this is a longer-term effort, it will ensure that the work is made readily available within the academic community in archival format. We also expect that the results from this work will inform a revised edition of our book on synthetic data [68].

As important, the results of this work will inform the broader future research agenda of EHIL. This means

[4] See <https://cippic.ca/> and <https://ccla.org/>

that the results will continue to have longer term impact and be part of other dissemination and knowledge translation efforts for multiple years, as well as informing technology development programs at EHIL.

10. Timeline and Monitoring

10.1. Project Timeline

Assuming a start date of April 1, 2021, we anticipate that this project will take one year to complete from start to finish. Figure 1 below outlines the expected duration for each of the components of this project. We structured the phases sequentially since earlier phases will inform the data collection and analysis in subsequent phases; however, work on each phase will at times overlap with work on a subsequent phase to allow for more efficient use of our time.

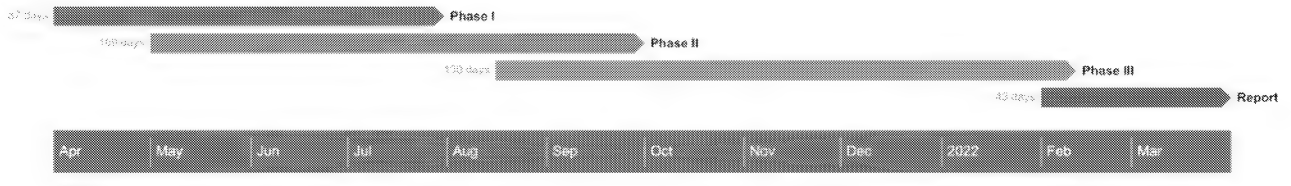


Figure 1: Timeline for Project Phases

Figure 2 below outlines the timeline for the environmental scan portion of the project. We project that this scan will take approximately four months to complete. We will begin by conducting our literature search and compiling article abstracts for review, which we estimate will take approximately 2 weeks. We have allotted approximately one month to the reviewing of abstracts to identify relevant articles. Another three weeks will be dedicated to obtaining and screening the full text articles. Finally, we have allotted approximately one month for reviewing the resulting set of full text articles.

In addition, we will support the legal review in Phase 2 by preparing summaries of information from the Environmental Scan that address the questions of "What are the privacy models that have been used to evaluate the privacy risks in synthetic data, and what are the results of empirical studies evaluating the privacy risks according to these models?" These summaries will be provided to our legal expert to inform the legal review, as well as being used to inform the final report. This information will be extracted during the review of the full text articles, and compiled afterwards. Therefore, we will add an additional week onto the end of Phase 2 for the compilation of these summaries.

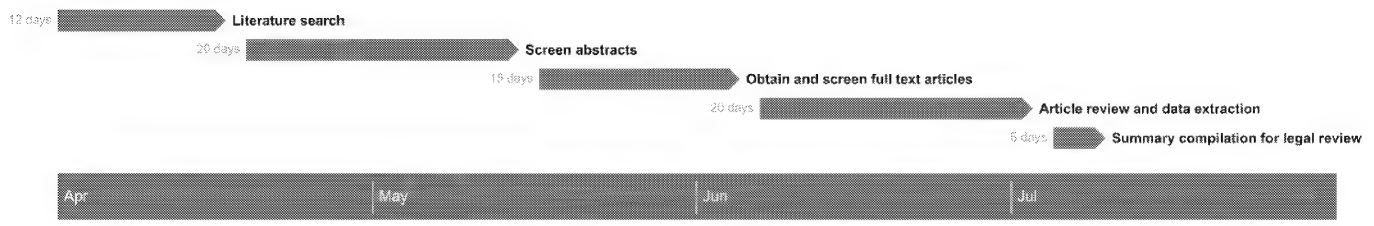


Figure 2: Timeline for Environmental Scan.

Our legal expert has estimated that the second portion of the project, the legal review, will take approximately 4 months to complete. The legal review can begin shortly after the project commences and can be conducted in parallel to the environmental scan. We have allotted an additional month to what our expert has estimated to allow for possible delays or other contingencies that may impact the timeline.

The third portion, interviews with privacy commissioners' staff, will be the most time consuming and we have allotted 6 months to complete this portion of the project. See Figure 3 below for a graphical illustration of the timeline. This portion of the project includes numerous sub-tasks beginning with the development of an interview guide in conjunction with stakeholders, which we are projecting will take approximately one month. Overlapping with this task will be the recruitment of participants which we have allotted seven weeks to complete, allowing for the sending of the initial email invitation as well as two reminder emails. Next, the scheduling of interviews will take place concurrently with recruitment as we plan to contact participants shortly after self-selection in order to schedule interviews. We have allotted three months for conducting the interviews, planning to schedule approximately two interviews per week and anticipating a two week blackout period for the holidays at the end of December/beginning of January. We are planning for transcription of the interview recordings and analysis of the data to take place in parallel to the interviews, beginning shortly after the interviews commence and ending three weeks after the interview period has concluded. Analysis of the interviews will include summarizing relevant information from the interviews (e.g., that related to the privacy risks posed by and regulation of synthetic data) that will be included in the results of the legal review.

Figure 3: Timeline for Interviews with Canadian Regulators

Finally, we have allotted approximately two months for the development, review and completion of the final report outlining the results of all three components of the project.

10.2. Dissemination of Results

Figure 4 below outlines the projected timeline for our dissemination and knowledge translation activities as described in Section 9.

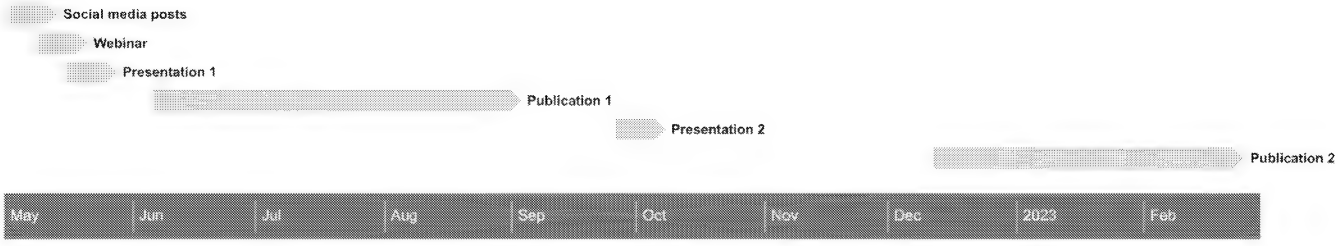


Figure 4: Timeline for dissemination of results / knowledge translation tasks.

Allowing one month for the release of the report by the OPC, we plan to begin our dissemination efforts in May 2022 with social media posts leading up to a webinar which outlines the results of the project. Although we do not know the exact dates of prospective conferences at this time, we are aware that the IAPP holds its Canadian Privacy Symposium in May each year and plan to submit a speaking proposal for that event based on the research undertaken for this project. During the summer of 2022, we plan to prepare a manuscript outlining the results of the environmental scan and/or legal review portion of the

project which will be submitted to an academic journal for publication. Again, although the exact dates are not yet known, IAPP holds several events in the fall months including their Data Protection Intensive and Privacy.Security.Risk conference. We plan to submit a speaking proposal to at least one of those events. The depicted two presentations would be the minimum number of presentations expected to be given on this subject. There is a good chance that other speaking opportunities will arise in 2022 and 23, and the number of presentations relating to this project would increase as a result. And finally, early in 2023 we plan to prepare a manuscript based on the interview portion of this project and/or follow up research conducted in this area that will be submitted to an academic journal for publication. As previously mentioned, the results of this project will inform future research at EHIL, and it is therefore expected to impact our dissemination and knowledge translation efforts for multiple years.

11. Budget

We include in our application for funding the form provided by the OPC in Schedule B—Eligible Costs. In this section, however, we provide a more detailed breakdown to explain these costs. These estimates include standard employer costs and benefits that are added to employee salaries, using effort estimates derived from the timeline described in the previous section where applicable.

The salaries to be funded by the Contributions program include those for a Research Coordinator and Data Analyst. The salary of the Principle Investigator will be provided as an in-kind contribution from the institution. A research coordinator will be needed to coordinate the project and the efforts of team members, compile a literature review on data synthesis, contact privacy commissioners and schedule interviews, and contribute to the drafting of reports. A data analyst will help analyze the interview data and draft the resulting report. The principle investigator will conduct interviews with privacy commissioners, aid in data analysis and be responsible for project deliverables.

In terms of contracted services, a legal expert will be required to conduct the legal analysis of data synthesis under PIPEDA and the CPPA. Anita Fineberg, who contributed to the previously funded OPC contributions project “E-Learning Course on Data Anonymization” will conduct the legal analysis and provide input in the creation of an interview guide. Transcription services will also be required to transcribe the audio recording of each interview in order to enable analysis of the data.

ELIGIBLE COSTS				OPC	IN-KIND	TOTAL
Salaries and Benefits						
Personnel	**FTE**	**Salary**	**Benefits**			
Research Scientist/Primary Investigator	0.25	30,000	7,500		$37,500	**$37,500**
Research Coordinator	0.15	8,951	2,238	$ 11,189		**$11,189**
Data analyst	0.12	8,400	2,100	$10,500		**$10,500**
Contractual Services						
Service		**Cost**	**HST**			
Analysis by Legal Expert		17,699	2,301	$ 20,000		**$20,000**
Transcription		1,125	146	$1,271		**$1,271**
Indirect Administrative Expenditures						
Indirect administrative expenditures (15%)				$6,444		**$6,444**
Total Budget				**$ 49,404**	**$37,500**	**$86,904**

12. Provincial/Territorial Support:

There is no specific provincial or territorial support for the project except for the in-kind support that would be provided by regulatory authorities participating in the interviews that are part of this project.

13. Acknowledgement of OPC Funding

The OPC's contributions will be acknowledged in all deliverables, articles, and presentations stemming from this work. The exact wording of the acknowledgement will follow the communications guidelines from the OPC using the OPC graphics and branding, as appropriate.

14. References

[1] European Medicines Agency, "External guidance on the implementation of the European Medicines

Agency policy on the publication of clinical data for medicinal products for human use," Sep. 2017.
[2] Health Canada, "Guidance document on Public Release of Clinical Information," Apr. 01, 2019. https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/profile-public-release-clinical-information-guidance.html.
[3] K. El Emam, *Guide to the De-Identification of Personal Health Information*. CRC Press (Auerbach), 2013.
[4] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the Crowd: The Privacy Bounds of Human Mobility," *Sci. Rep.*, vol. 3, Mar. 2013, doi: 10.1038/srep01376.
[5] Y.-A. de Montjoye, L. Radaelli, V. K. Singh, and A. "Sandy" Pentland, "Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata," *Science*, vol. 347, no. 6221, pp. 536–539, Jan. 2015, doi: 10.1126/science.1256297.
[6] L. Sweeney, J. Su Yoo, L. Perovich, K. E. Boronow, P. Brown, and J. Green Brody, "Re-identification Risks in HIPAA Safe Harbor Data: A study of data from one environmental health study," *J. Technol. Sci.*, no. 2017082801, Aug. 2017, Accessed: Mar. 23, 2020. [Online]. Available: https://techscience.org/a/2017082801/.
[7] J. Su Yoo, A. Thaler, L. Sweeney, and J. Zang, "Risks to Patient Privacy: A Re-identification of Patients in Maine and Vermont Statewide Hospital Data," *J. Technol. Sci.*, no. 2018100901, Oct. 2018, Accessed: Mar. 23, 2020. [Online]. Available: https://techscience.org/a/2018100901/.
[8] L. Sweeney, "Matching Known Patients to Health Records in Washington State Data," Harvard University. Data Privacy Lab, 2013.
[9] L. Sweeney, M. von Loewenfeldt, and M. Perry, "Saying it's Anonymous Doesn't Make It So: Re-identifications of 'anonymized' law school data," *J. Technol. Sci.*, no. 2018111301, Nov. 2018, Accessed: Mar. 23, 2020. [Online]. Available: https://techscience.org/a/2018111301/.
[10] A. Zewe, "Imperiled information: Students find website data leaks pose greater risks than most people realize," *Harvard John A. Paulson School of Engineering and Applied Sciences*, Jan. 17, 2020. https://www.seas.harvard.edu/news/2020/01/imperiled-information (accessed Mar. 23, 2020).
[11] K. Bode, "Researchers Find 'Anonymized' Data Is Even Less Anonymous Than We Thought," *Motherboard: Tech by Vice*, Feb. 03, 2020. https://www.vice.com/en_ca/article/dygy8k/researchers-find-anonymized-data-is-even-less-anonymous-than-we-thought (accessed May 11, 2020).
[12] E. Clemons, "Online Profiling and Invasion of Privacy: The Myth of Anonymization," *HuffPost*, Feb. 20, 2013.
[13] C. Jee, "You're very easy to track down, even when your data has been anonymized," *MIT Technology Review*, Jul. 23, 2019. https://www.technologyreview.com/2019/07/23/134090/youre-very-easy-to-track-down-even-when-your-data-has-been-anonymized/ (accessed May 11, 2020).
[14] G. Kolata, "Your Data Were 'Anonymized'? These Scientists Can Still Identify You," *The New York Times*, Jul. 23, 2019.
[15] N. Lomas, "Researchers spotlight the lie of 'anonymous' data," *TechCrunch*, Jul. 24, 2019. https://techcrunch.com/2019/07/24/researchers-spotlight-the-lie-of-anonymous-data/ (accessed May 11, 2020).
[16] S. Mitchell, "Study finds HIPAA protected data still at risks," *Harvard Gazette*, Mar. 08, 2019. https://news.harvard.edu/gazette/story/newsplus/study-finds-hipaa-protected-data-still-at-risks/ (accessed May 11, 2020).
[17] S. A. Thompson and C. Warzel, "Twelve Million Phones, One Dataset, Zero Privacy," *The New York Times*, Dec. 19, 2019.

[18] "'Anonymised' data can never be totally anonymous, says study," *the Guardian*, Jul. 23, 2019.
[19] R. Rocha, "The data-driven pandemic: Information sharing with COVID-19 is 'unprecedented,'" *CBC News*, Canada, Mar. 17, 2020.
[20] "ViriHealth – Canada's Coronavirus COVID-19 Tracker," *ViriHealth*. https://virihealth.com/ (accessed Apr. 09, 2020).
[21] "Coronavirus COVID-19 (2019-nCoV)." https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6 (accessed Apr. 09, 2020).
[22] "Coronavirus COVID-19 Global Cases by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)," *Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)*, 2020. https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6.
[23] "Novel Coronavirus (COVID-19)," *HealthMap*, 2020. https://www.healthmap.org/covid-19/ (accessed Apr. 09, 2020).
[24] "Tracking the coronavirus," *CBC News*, 2020. https://newsinteractives.cbc.ca/coronavirustracker/ (accessed Apr. 09, 2020).
[25] "Genomic epidemiology of novel coronavirus - Global subsampling," *Nextstrain*, 2020. https://nextstrain.org/ncov/global (accessed Apr. 09, 2020).
[26] S. Layne, J. Hyman, D. Morens, and J. Taubenberger, "New coronavirus outbreak: Framing questions for pandemic prevention," *Sci. Transl. Med.*, vol. 12, no. 534, Mar. 2020, doi: 10.1126/scitranslmed.abb1469.
[27] M. Downey, "Sharing data and research in a time of global pandemic," *Duke University Libraries*, Mar. 17, 2020. https://blogs.library.duke.edu/bitstreams/2020/03/17/sharing-data-and-research-in-a-time-of-global-pandemic/ (accessed Apr. 08, 2020).
[28] A. Ng, "Coronavirus pandemic changes how your privacy is protected," *CNET*, Mar. 21, 2020. https://www.cnet.com/news/coronavirus-pandemic-changes-how-your-privacy-is-protected/ (accessed Apr. 08, 2020).
[29] A. L. Beam and I. S. Kohane, "Big Data and Machine Learning in Health Care," *JAMA*, vol. 319, no. 13, pp. 1317–1318, Apr. 2018, doi: 10.1001/jama.2017.18391.
[30] Z. Wang, P. Myles, and A. Tucker, "Generating and Evaluating Synthetic UK Primary Care Data: Preserving Data Utility Patient Privacy," in *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, Jun. 2019, pp. 126–131, doi: 10.1109/CBMS.2019.00036.
[31] "Synthetic data at CPRD." https://www.cprd.com/content/synthetic-data (accessed Sep. 24, 2020).
[32] N3C, "Synthetic Data Workstream | N3C." https://covid.cd2h.org/N3C_synthetic_data (accessed Sep. 24, 2020).
[33] "National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment | Journal of the American Medical Informatics Association | Oxford Academic." https://academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocaa196/5893482?login=true (accessed Jan. 16, 2021).
[34] "#opendata4covid19 Website User Manual." Ministry of Health and Welfare; Health Insurance Review & Assessment Service (HIRA), Apr. 2020, Accessed: Apr. 08, 2020. [Online]. Available: https://rtrod-assets.s3.ap-northeast-2.amazonaws.com/static/tools/manual/COVID-19+website+manual_v2.1.pdf.
[35] "Coronavirus privacy: Are South Korea's alerts too revealing?"

[36] Arej Dajani *et al.*, "The modernization of statistical disclosure limitation at the U.S. Census Bureau," Census Scientific Advisory Committee Meeting, 2017.
[37] Government Accountability Office, "Artificial Intelligence: Emerging Opportunities, challenges, and Implications," Mar. 2018.
[38] McKinsey Global Institute, "Artificial Intelligence: The Next Digital Frontier ?," Jun. 2017.
[39] Deloitte Insights, "State of AI in the Enterprise, 2nd Edition," 2018.
[40] B. Lorica and P. Nathan, "The State of Machine Learning Adoption in the Enterprise," O'Reilly, 2018.
[41] Office of the Privacy Commissioner of Canada, "A Regulatory Framework for AI: Recommendations for PIPEDA Reform." Nov. 12, 2020, Accessed: Feb. 02, 2021. [Online]. Available: https://www.priv.gc.ca/en/about-the-opc/what-we-do/consultations/completed-consultations/consultation-ai/reg-fw_202011/.
[42] R. Little, "Statistical Analysis of Masked Data," *J. Off. Stat.*, vol. 9, no. 2, pp. 407–426, 1993.
[43] D. Rubin, "Discussion: Statistical Disclosure Limitation," *J. Off. Stat.*, vol. 9, no. 2, pp. 462–468, 1993.
[44] J. Reiter, "Using CART to generate partially synthetic, public use microdata," *J. Off. Stat.*, vol. 21, no. 3, pp. 441–462, 2005.
[45] Z. Wan, Y. Zhang, and H. He, "Variational autoencoder based synthetic data generation for imbalanced learning," in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, Nov. 2017, pp. 1–7, doi: 10.1109/SSCI.2017.8285168.
[46] L Gootjes-Dreesbach, M Sood, A Sahay, and M Hofmann-Apitius, "Variational Autoencoder Modular Bayesian Networks (VAMBN) for Simulation of Heterogeneous Clinical Study Data - Abstract - Europe PMC." https://europepmc.org/article/ppr/ppr91638 (accessed Jan. 06, 2020).
[47] Z. Zhang, C. Yan, D. A. Mesa, J. Sun, and B. A. Malin, "Ensuring electronic medical record simulation through better training, modeling, and evaluation," *J. Am. Med. Inform. Assoc.*, doi: 10.1093/jamia/ocz161.
[48] J. Hu, "Bayesian Estimation of Attribute and Identification Disclosure Risks in Synthetic Data," *ArXiv180402784 Stat*, Apr. 2018, Accessed: Mar. 15, 2019. [Online]. Available: http://arxiv.org/abs/1804.02784.
[49] J. P. Reiter, "New Approaches to Data Dissemination: A Glimpse into the Future (?)," *CHANCE*, vol. 17, no. 3, pp. 11–15, Jun. 2004, doi: 10.1080/09332480.2004.10554907.
[50] J. Taub, M. Elliot, M. Pampaka, and D. Smith, "Differential Correct Attribution Probability for Synthetic Data: An Exploration," in *Privacy in Statistical Databases*, 2018, pp. 122–137.
[51] J. Hu, J. P. Reiter, and Q. Wang, "Disclosure Risk Evaluation for Fully Synthetic Categorical Data," in *Privacy in Statistical Databases*, 2014, pp. 185–199.
[52] L. Wei and J. P. Reiter, "Releasing synthetic magnitude microdata constrained to fixed marginal totals," *Stat. J. IAOS*, vol. 32, no. 1, pp. 93–108, Jan. 2016, doi: 10.3233/SJI-160959.
[53] N. Ruiz, K. Muralidhar, and J. Domingo-Ferrer, "On the Privacy Guarantees of Synthetic Data: A Reassessment from the Maximum-Knowledge Attacker Perspective," in *Privacy in Statistical Databases*, 2018, pp. 59–74.
[54] J. P. Reiter, "Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study," *J. R. Stat. Soc. Ser. A Stat. Soc.*, vol. 168, no. 1, pp. 185–205, 2005, doi: 10.1111/j.1467-985X.2004.00343.x.
[55] J. Drechsler and J. P. Reiter, "Accounting for Intruder Uncertainty Due to Sampling When Estimating Identification Disclosure Risks in Partially Synthetic Data," in *Privacy in Statistical Databases*, 2008, pp. 227–238.
[56] J. Drechsler and J. P. Reiter, "An empirical evaluation of easily implemented, nonparametric

methods for generating synthetic datasets," *Comput. Stat. Data Anal.*, vol. 55, no. 12, pp. 3232–3243, Dec. 2011, doi: 10.1016/j.csda.2011.06.006.

[57] J. P. Reiter and R. Mitra, "Estimating Risks of Identification Disclosure in Partially Synthetic Data," *J. Priv. Confidentiality*, vol. 1, no. 1, Apr. 2009, doi: 10.29012/jpc.v1i1.567.

[58] A. Dandekar, R. Zen, and S. Bressan, "A comparative study of synthetic dataset generation techniques," National University of Singapore, TRA6/18, 2018. [Online]. Available: https://dl.comp.nus.edu.sg/bitstream/handle/1900.100/7050/TRA6-18.pdf?sequence=1&isAllowed=y.

[59] A. Guo, R. E. Foraker, R. M. MacGregor, F. M. Masood, B. P. Cupps, and M. K. Pasque, "The Use of Synthetic Electronic Health Record Data and Deep Learning to Improve Timing of High-Risk Heart Failure Surgical Intervention by Predicting Proximity to Catastrophic Decompensation," *Front. Digit. Health*, vol. 2, 2020, doi: 10.3389/fdgth.2020.576945.

[60] N. Nama *et al.*, "Crowdsourcing the Citation Screening Process for Systematic Reviews: Validation Study," *J. Med. Internet Res.*, vol. 21, no. 4, p. e12953, 2019, doi: 10.2196/12953.

[61] J. Creswell, *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage, 2003.

[62] J. Creswell and V. Plano-Clark, *Designing and conducting mixed methods research*. Sage, 2007.

[63] G. Gaskell, "Individual and Group Interviewing," in *Qualitative Researching with Text, Image and Sound: A practical handbook*, London: SAGE Publications, 2000, pp. 38–56.

[64] J. Creswell, *Qualitative Inquiry and Research Design: Choosing Among Five Approaches*. Thousand Oaks, CA: SAGE Publications, 1998.

[65] M. Patton, *Qualitative evaluation and research methods*. Sage Publications, 1990.

[66] A. Strauss and J. Corbin, *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage, 1998.

[67] B. G. Glaser and A. L. Strauss, *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine, 1967.

[68] K. El Emam, L. Mosquera, and R. Hoptroff, *Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data*. O'Reilly, 2020.

Appendix: Literature Search Strategy

We conducted a literature search on January 18 and 19, 2021 to obtain publications related to data synthesis and synthetic data that were published in the year 2020. We searched the Scopus database which indexes content from 24,600 active titles and 5,000 publishers (including general science, medical informatics and computer science journals).

Search Terms

Core Concepts

We searched in *All Fields* for the Core Concepts *"synthetic data" OR "data synthesis"*, limiting the results to articles published in 2020 only. Including *"data synthesis"* in the Scopus search returned a good deal of systematic review papers and meta-analyses, so that term was dropped from the core concepts and only *"synthetic data"* was used. Also, to further limit the return of systematic reviews and meta-analyses, the Boolean operator *NOT* was added with the terms *"systematic review" OR "meta-analysis" OR "knowledge synthesis" OR "research synthesis"*.

Methods of Synthetic Data Generation

To search for specific methods of synthetic data generation, to the Core Concepts above were added the Boolean operator *AND* plus the name of the specific method. For example, to search for GANs *AND "generative adversarial network*"* was added. These searches were conducted for GANs, Bayesian networks and Copula methods.

Search Term Examples

Below are some examples of the search terms used in Scopus. Note, in Scopus not specifying a field code returns results for ALL field codes (equivalent to searching all fields)[5].

Core Concepts search

"synthetic data" AND NOT ("systematic review" OR "meta-analysis" OR "knowledge synthesis" OR "research synthesis") AND (LIMIT-TO (PUBYEAR , 2020))

GANs Search

("synthetic data" AND NOT ("systematic review" OR "meta-analysis" OR "knowledge synthesis" OR "research synthesis")) AND " generative adversarial network" AND (LIMIT-TO (PUBYEAR , 2020))*

Bayesian networks search

("synthetic data" AND NOT ("systematic review" OR "meta-analysis" OR "knowledge

[5] See: https://service-elsevier-com.proxy.bib.uottawa.ca/app/answers/detail/a_id/11236/supporthub/scopus/session/L2F2LzEvdGltZS8xNjExMDY5NTc1L2dlbi8xNjExMDY5NTc1L3NpZC9mVVliTElNTnhCYXJQUFVTOXFrVmhhdVhtR3A1TFF4VVpDJTdFZlkyUWZVSWtXaUhaTU5RUXhMbkg4dDFPdWY3T0lhaHNDcG9vWnY3SDFZSVpCSG5FTFJ6akFDOGtzQVNxUXZUYWk2dFN3RmlYJTdFWW9LbCU3RVJUVFhwX2clMjElMjE%3D/

synthesis" OR "research synthesis")) AND "Bayesian network" AND (LIMIT-TO (PUBYEAR , 2020))*

Copula search

("synthetic data" AND NOT ("systematic review" OR "meta-analysis" OR "knowledge synthesis" OR "research synthesis")) AND "Copula" AND (LIMIT-TO (PUBYEAR , 2020))

CHEO

uOttawa

Interview Study of Canadian Privacy Regulators on Regulating Synthetic Data:

Background Information

Khaled El Emam

19th January 2022

401 Smyth Road, Ottawa, ON K1H 8L1, Canada
Tel: (613) 737-7600 www.cheo.on.ca

401, chemin Smyth, Ottawa (ON) K1H 8L1, Canada
Tél. : (613) 737-7600 www.cheo.on.ca

Background

Interest in synthetic data has been growing quite rapidly over the last few years. This has been driven by three simultaneous trends. The first is the demand for large amounts of data to train and build artificial intelligence and machine learning (AIML) models. The second is recent work that has demonstrated effective methods to generate high quality synthetic data. Third, trust in contemporary de-identification methods has been eroded by repeated claims of successful re-identification attacks on anonymized data, reducing confidence in contemporary practices. This has created a need, and a demand, for more advanced privacy enhancing technologies (PETs), such as synthetic data generation.

To create synthetic datasets, one first gets a real dataset and trains an AIML model on it. This is called a **generative model**. Then new data are generated from the trained model. This new data is the synthetic data. Because this new data is generated from a model, there is no one-to-one mapping between the synthetic records and real people. If constructed properly, the synthetic data are effectively fake data but that retain the statistical properties of the original real data. The concept of "re-identification" for synthetic data does not really make sense since the synthetic records are not mapped to real people.

Synthetic datasets can still have some privacy risks, but these are not re-identification risks under the common definition of the term. The main privacy risks that have been examined are: attribution risk and membership disclosure risk. Attribution risk is where a record that looks similar to a real person exists in the synthetic data and an adversary learns something new about that person from that similar synthetic record. Membership disclosure means that the adversary can learn that a person's record was in the real dataset used to generate the synthetic data. The evidence that has accumulated thus far suggests that these privacy risks have tended to be quite low in practice.

Uses of synthetic data are growing quite fast as it is increasingly seen as a safe way to create useful and non-identifiable data. For example, large health data custodians are creating synthetic variants of their data holdings and making these available to a broad community of analysts. There has also been growing adoption in the financial services sector. Perhaps one of the most public examples is the adoption of synthetic data by the US Census Bureau. In addition to that, large sums of money are being invested in a multitude of synthetic data generation startups that are developing software to make this technology more easily available.

As the adoption of synthetic data increases in practice, whether it should be regulated, and if so, the manner in which it is regulated, become important questions. The timeliness of this issue is exemplified by a new report by the Expert Advisory Group on the Pan-Canadian Health Data Strategy which noted that a certain "privacy chill" arising from risk-averse interpretations of health data sharing rules is hurting patient care and hampering responses to health crises.

The purpose of this series of interviews is to explore how Canadian regulators think about and how they may approach synthetic data. It is part of a project funded by the Office of the Privacy Commissioner of Canada under their Contributions Program.

As background reading, we have included part of a chapter on privacy risks in synthetic data from the perspective of the GDPR, HIPAA, and CCPA.[1] Although that analysis is not specific to Canada, it starts to address some of the relevant regulatory questions.

Interview Questions

The following are the questions which we will use to guide the interview. We will deviate from this list if the conversation leads us in new directions.

1. Have you had experience with synthetic data in your jurisdiction? For example, entities asking your office for advice on generating or using synthetic data, or synthetic data being a part of investigations?

2. Training a generative model is a form of data use or processing. Does this use require additional specific consent from data subjects or would the fact that it is a privacy protective technology that enhances the rights of the data subjects mitigate against that? Despite ambiguity in some statutes across the country, current practice thus far has been to treat the creation of non-identifiable information as a form of processing that does not require additional consent.

3. Should synthetic data be regulated under privacy laws, and if so, how? The implications of regulating fake data could arguably be quite impactful unless only certain types of synthetic data are regulated. How would we define these carve-outs, if any?

4. Is a data custodian able to delegate the creation of synthetic data to a third party (a sub-contractor)? What conditions would apply under these circumstances?

5. Given the concerns with non-identifiable data that have been expressed in the media and by regulators recently, do you think there is a need for guidance or standards on the generation of synthetic data? How would there be assurance that these standards are being applied properly? Do we need guidance on the ethical uses of synthetic data?

6. There is growing evidence that machine learning models can be attacked to recover part or all of the training data. This has resulted in some saying that machine learning models trained on personally identifying data should be treated as personal information when they are used and disclosed. However, because synthetic data is not identifiable then machine learning models trained on synthetic data would not need to be treated as personal information. What are your thoughts on this perspective?

7. Do data custodians need to inform data subjects if they use their data to create synthetic datasets (this would be a form of transparency rather than consent)?

8. Can synthetic data be disclosed to anyone given that it is not identifiable information?

9. Can synthetic data be used for any purpose given that it is not identifiable information?

[1] This analysis was performed by Mike Hintze from Hintze Law PLLC.

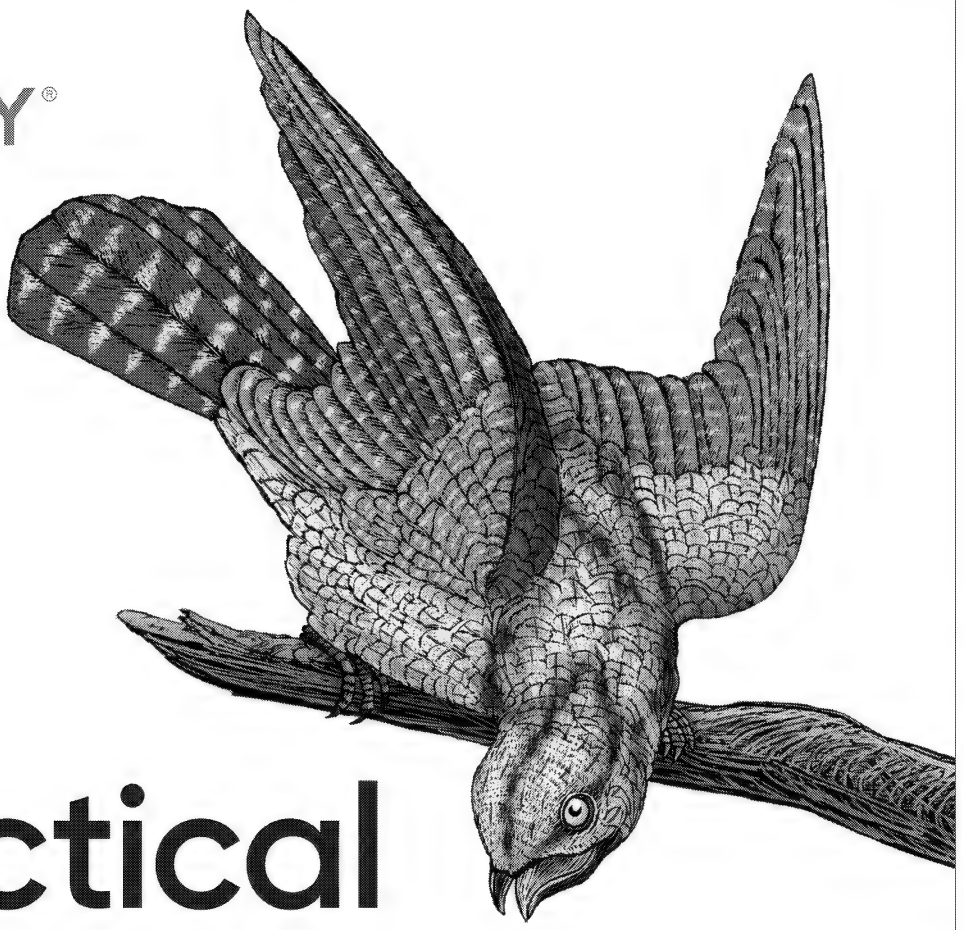
Interview Protocol

The protocol for this interview will be as follows:

- The interview will be conducted by Khaled El Emam from the University of Ottawa.
- The interview is expected to last 45 minutes.
- The interview will be recorded and transcribed. This will ensure that important points are not missed. The recordings will be deleted after the study. The transcript will be kept for seven years in case questions arise about publications from this work.
- The report and any subsequent publications will not attribute any comment to an identifiable interviewee. All reported information will be of trends and aggregate information.
- We will provide you with a copy of the final report which we submit to the federal commissioner's office once it is completed. If you would like to review a draft please let me know. However, we expect the turn-around times for the review cycle to be quite tight.

Thank you very much for your time. It is greatly appreciated.

O'REILLY®



Practical Synthetic Data Generation

Balancing Privacy and the Broad Availability of Data

Khaled El Emam,
Lucy Mosquera &
Richard Hoptroff

In our example from the previous chapter with using decision trees for synthesis on the hospital discharge data, we found that 4% of the records were unique in the synthetic data and were also unique in the real dataset. Therefore, these records can be removed from the synthetic dataset as a privacy protection measure.

This approach is really quite conservative and can be considered a simple first step to empirically evaluating the identity disclosure risks in a synthetic dataset. More sophisticated methods can be applied to statistically estimate the probability of matching a synthetic record to a real person, accounting for different attack methods that an adversary can use.

How Privacy Law Impacts the Creation and Use of Synthetic Data

Synthetic data offers a compelling solution to data sharing and data-access barriers—one that promotes greater scientific and commercial research while protecting individual privacy.[6]

An original set of real personal information is used in the creation and evaluation of a synthetic dataset. A synthetic dataset is generated from a real dataset. The synthetic dataset has the same statistical properties as the real data. But it is not real data. It is not data about or related to any real individual person or people. A single record in a synthetic dataset does not correspond to an individual or record in the real dataset. And to ensure that the resulting synthetic dataset does not inadvertently reveal information about a real person from the original dataset, a privacy assurance process evaluates the privacy risk of the synthetic data—comparing the real and the synthetic data to assess and remove any such risk.

Synthetic data differs from what is traditionally thought of as the de-identification of data. De-identification is a means of altering a dataset to remove, mask, or transform direct and indirect identifiers. But the de-identified data is still real data related to real individuals. It has just made it less likely that any individual in the record can be identified from the data. Depending on the method and strength of the de-identification, it can be an excellent risk-mitigation measure. But depending on the applicable laws, it may still be treated as personal information, and there still can be significant regulatory overhead. Contracts with data recipients may need to be in place, security precautions must be taken, and distribution may need to be limited.

6 This section of the chapter is for informational purposes only and is not intended to provide, nor shall it be construed as providing, any any legal opinion or conclusion, does not constitute legal advice, and is not a substitute for obtaining professional legal counsel from a qualified attorney on your specific matter. The material here was prepared by Mike Hintze from the firm Hintze Law.

Synthetic data is different. It is not real data related to real people. There is no link between a synthetic dataset and records in the original (real) dataset. If done properly, the creation of synthetic data should result in a dataset that cannot be reverse engineered to reveal identities of real people or information specific to a real person.[7] For any given synthetic dataset, this conclusion is testable and verifiable through statistical analysis. Thus, a properly created and verified synthetic dataset that is not constrained by privacy law can be freely distributed (including publicly released) and used broadly for analysis and research.

But that does not mean that privacy laws are irrelevant. Because synthetic data must start with a real dataset, the handling and use of that real dataset is still likely to be regulated by privacy law.

If an organization does not have the capability and expertise to create synthetic data in-house, it may need to share the original (real) dataset with a service provider to create the synthetic data. That sharing is also likely to be subject to privacy law.

This section addresses how the creation and use of synthetic data is regulated under three key privacy laws: the European General Data Protection Regulation (GDPR),[8] the California Consumer Privacy Act (CCPA),[9] and the US Health Insurance Portability and Accountability Act (HIPAA).[10]

For each of these privacy laws, this chapter will examine three key questions:

- Is the use of the original (real) dataset to generate and/or evaluate a synthetic dataset restricted or regulated under the law?
- Is sharing the original dataset with a third-party service provider to generate the synthetic dataset restricted or regulated under the law?
- Does the law regulate or otherwise affect (if at all) the resulting synthetic dataset?

In sum, while these laws regulate or potentially regulate the generation and evaluation of synthetic data, as well as the sharing of the original dataset with third-party service

7 This conclusion holds true even if the person using the synthetic dataset has or could gain access to the original dataset. That would not occur in most cases, since the key objective of creating synthetic data is to enable the benefits of data use and analysis without giving access to real, personal data. Nevertheless, it may be worthwhile to have in place additional safeguards such as strong access controls on the original dataset, and contractual prohibitions on any attempts to reverse engineer or link the synthetic data to the original data.

8 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119) 1 (hereinafter "GDPR").

9 California Consumer Privacy Act of 2018, Cal. Civ. Code §§1798.100-1798.199 (hereinafter "CCPA").

10 Health Insurance Portability and Accountability Act of 1996, Pub. L. 104-191 (hereinafter "HIPAA"); Standards for Privacy of Individually Identifiable Health Information, 45 C.F.R. Parts 160 and 164 (hereinafter "HIPAA Privacy Rule").

providers, none pose a significant barrier to doing so. Sharing the original data with a service provider is permitted as long as an appropriate contract is in place and the parties adhere to its requirements. And once a fully synthetic dataset is created, this data should be seen as falling outside the scope of these laws, and therefore not subject to any restrictions on the subsequent use or dissemination of the data (including making the data publicly available).

We conclude the section with an analysis of an opinion on what makes information identifiable, published by an advisory body of European regulators (the Article 29 Working Party). We provide a pragmatic interpretation of that opinion and explain how that can be applied to synthetic data.

Issues Under the GDPR

Here we address some common questions regarding how the GDPR applies to synthetic data generation and use.

Is the use of the original (real) dataset to generate and/or evaluate a synthetic dataset restricted or regulated under the GDPR?

Yes. The GDPR regulates any "processing" of personal data. And "processing" is defined as "any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means."[11] Because the generation of synthetic data involves the processing of a real dataset, the obligations that the GDPR imposes on the processing of personal data apply to this operation.

In particular, the GDPR requires there to be a "legal basis" to process personal data. Thus, to the extent that the original dataset includes personal data, the use of that dataset to generate or evaluate a synthetic dataset requires a legal basis. There are several legal bases available under the GDPR. One well-known legal basis is the consent of the individual.

But obtaining consent from every individual contained in a dataset in order to develop a synthetic dataset will often be impractical or impossible. Further, seeking consent from data subjects to process data in order to create synthetic data (and excluding the data from those individuals who do not consent) could undermine the statistical validity of the generated synthetic data because there is significant evidence of consent bias.[12]

11 GDPR Art. 4(2).

12 See Khaled El Emam et al., "A Review of Evidence on Consent Bias in Research," *The American Journal of Bioethics* 13, no. 4 (2013): 42–44. *https://oreil.ly/5x5kg*; Michelle E. Kho et al., "Written Informed Consent and Selection Bias in Observational Studies Using Medical Records: Systematic Review," *BMJ* 338:b866 (March 2009). *https://doi.org/10.1136/bmj.b866*.

Instead, a more practical and appropriate legal basis will be "legitimate interests." This legal basis applies when the legitimate interests of the data controller or a third party outweigh the interests or rights of the data subject. Inherent in the use of this legal basis is a balancing test. In this context, one must consider the interest in processing personal data in order to create a synthetic dataset and weigh that interest against the risks to the data subject.

An organization that has a need or desire to use data for a purpose that a synthetic dataset can help achieve, or that wishes to advance beneficial research while reducing the organization's legal risk and protecting the privacy of individuals, will have a very strong interest in the creation of a synthetic dataset that can be used for research in lieu of using real data. On the other side of the equation, assuming the creation of synthetic data is done in a secure environment, there is little or no risk to the data subject. On the contrary, the data subject has an interest in the creation of the synthetic data because it eliminates the risk inherent in sharing and using the original (real) dataset for a research purpose when the synthetic data can be used instead. Thus, the legitimate interests balancing test comes out strongly in favor of using the personal data to create the synthetic dataset.

Beyond the need to establish a legal basis for processing, the GDPR includes a number of additional obligations relating to the collection, use, and disclosure of personal data—which apply in this scenario just as they apply to any processing of personal data. Thus, the organization handling the original dataset must ensure that the personal data is kept secure and protected from unauthorized access or disclosure.[13] The organization must meet its notice and transparency obligations, so it may be prudent to ensure that the applicable privacy notice(s) contemplate and disclose the types of processing that are involved in the creation and testing of synthetic datasets.[14] And the organization must maintain records of its processing activities; here too the organization should simply make sure that this use of data to create synthetic data is in some way reflected in those records.

However, these are obligations the organization must meet with respect to its collection and processing of the original dataset in any event, whether or not that set is used in the generation of synthetic data. The use of personal data to create synthetic data will, at most, have a modest impact on how the organization meets those

13 GDPR Art. 32 ("Taking into account the state of the art, the costs of implementation and the nature, scope, context and purposes of processing as well as the risk of varying likelihood and severity for the rights and freedoms of natural persons, the controller and the processor shall implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk").

14 GDPR Art. 13(1) ("the controller shall, at the time when personal data are obtained, provide the data subject with all of the following information… (c) the purposes of the processing for which the personal data are intended").

obligations. But it does not create fundamentally new obligations, nor does it significantly increase the burden or difficulty of meeting these existing obligations.

Is sharing the original dataset with a third-party service provider to generate the synthetic dataset restricted or regulated under the GDPR?

Under the GDPR, any entity processing personal data will be either a "data controller" or a "data processor." A data controller is an entity that "alone or jointly with others, determines the purposes and means of the processing of personal data." A data processor is an entity that processes personal data on behalf of, and at the direction of, the controller. For the purposes of this discussion, we can assume that the owner of the dataset is the data controller, and the service provider that the controller hires to generate synthetic data from that original dataset is a data processor.

A data controller can provide personal data to a data processor as necessary to enable the data processor to perform a service at the direction of and on behalf of the data controller. So, sharing an original dataset with a third-party service provider to generate a synthetic dataset is permitted under the GDPR. However, the GDPR imposes certain restrictions on that data sharing and on the parties involved.

A controller that wishes to share personal data with a processor has a duty of care in selecting a processor that can provide "sufficient guarantees" that it will process personal data in compliance with the requirements of the GDPR and will protect the rights of the data subject(s).

The GDPR further requires that there be a contract between the controller and the processor that obligates the processor to do the following:

- Process the personal data "only on documented instructions from the controller"
- Ensure that each person processing the personal data is subject to a duty of confidentiality with respect to the data
- Implement and maintain reasonable security procedures and practices to protect personal data
- Engage a subcontractor only pursuant to a written contract that passes through the data protection requirements and only with the general or specific prior consent of the controller[15]

15 GDPR Art. 28(2), (3)(d), and (4). "General" consent for the processor to use subcontractors can be provided in advance, including as part of the contract, so long as the processor informs the controller of any addition of replacement of subcontractors, and gives the controller the opportunity to object.

- Assist the controller in enabling data subjects to exercise their rights under the GDPR[16]
- Assist the controller as needed to meet the controller's obligations with respect to data security, notification of data breaches, risk assessments, and consultation with regulators
- Delete or return all personal data to the controller at the completion of the service(s)—unless retention is required by law
- Provide to the controller, upon request, "all information necessary to demonstrate compliance with [the processor's] obligations…and allow for and contribute to audits, including inspections, conducted by the controller or another auditor mandated by the controller"

As long as the contract with the required terms for data processors is in place, and those measures are adhered to, providing the original dataset to a service provider in order to create a synthetic dataset will be permissible under the GDPR.

Does the GDPR regulate or otherwise affect (if at all) the resulting synthetic dataset?

Once the synthetic dataset has been generated, any regulation of that dataset under the GDPR depends on whether it can be considered "personal data."

The GDPR defines "personal data" as follows:

> Any information relating to an identified or identifiable natural person ("data subject"); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.[17]

Synthetic data is not real data about a person. Although it is based on a real dataset, a single record in a synthetic dataset does not correspond to an individual or record in the original (real) dataset. Thus, a record in a synthetic dataset does not relate to an actual natural person. It does not include an identifier that corresponds to an actual natural person. It does not reference the physical, physiological, genetic, mental, economic, cultural, or social identity of an actual natural person. In short, a fully synthetic dataset does not meet the GDPR definition of "personal data." As such, it is

16 GDPR Art. 28(3)(e). In cases where a data processor holds personal data for a relatively short period of time, as would be the case here, where the original dataset containing personal data is processed by the service provider for only as long as is required to create and test the synthetic dataset, it is unlikely that this obligation to assist the data controller with requests from data subjects (such as requests to access or delete data) would apply in a significant way.

17 GDPR Art. 4(1).

outside the scope of the GDPR. It therefore can be used and distributed, including being made publicly available, without restriction.

Issues Under the CCPA

Here we address some common questions regarding how the CCPA applies to synthetic data generation and use.

Is the use of the original (real) dataset to generate and/or evaluate a synthetic dataset restricted or regulated under the CCPA?

Unlike the GDPR, the CCPA does not require the establishment of a legal basis for the processing of personal information. Nor does it place significant restrictions on a company's collection or internal use of personal information. Instead, it is largely focused on regulating the "sales" of personal information, which is defined broadly to cover many transfers of personal information in a commercial context.

As a result, the act of using an existing dataset to create a synthetic dataset is not specifically regulated by the CCPA. Thus, the CCPA does not prevent or restrict the use of personal information to generate and/or evaluate a synthetic dataset.

Instead, as with the GDPR, such data use may be subject to some CCPA obligations, such as providing notice of how the personal information is used, which will apply to the organization whether or not it uses the data to generate synthetic data.

Is sharing the original dataset with a third-party service provider to generate the synthetic dataset restricted or regulated under the CCPA?

As noted previously, the CCPA regulates the "sale" of personal information, and sales are defined very broadly. However, certain transfers of personal information to a "service provider" are exempt from the definition of "sale."[18] Specifically, if personal data is transferred by a business to a service provider, that transfer will not be regulated as a sale under the CCPA as long as the following requirements are met:

- The business has provided notice that personal information will be shared with service providers
- The service provider does not collect, use, sell, or disclose the personal data for any purpose other than as necessary to provide the service(s) on behalf of the business
- There is a written contract between the business and the service provider that specifies the service provider is prohibited from retaining, using, selling, or

18 Note that virtually every organization shares some data with a service provider from time to time, so the organization's privacy notice should already have such a disclosure.

disclosing the personal information for any purpose other than performing the services specified in the contract on behalf of the business

Thus, as long as these criteria are met, including having a contract in place between the business and the service provider,[19] a business subject to the CCPA can share a dataset containing personal information with a service provider that uses it to generate synthetic data on behalf of that business.

Does the CCPA regulate or otherwise affect (if at all) the resulting synthetic dataset?

The CCPA defines "personal information" as any information "that identifies, relates to, describes, is reasonably capable of being associated with, or could reasonably be linked, directly or indirectly, with a particular consumer or household." While this is a very broad definition of personal information, it should not include synthetic data. As noted previously, synthetic data is not real data that relates to a real person. When a synthetic dataset is created using a real dataset, there is no association between an individual record in a real set and an individual record in the resulting synthetic dataset. Thus, records in a synthetic set should not be seen as being associated, linked, or related to a particular real consumer or household.

Further, the CCPA definition of personal information specifies that it does not include aggregate consumer information. And "aggregate consumer information" is defined as "information that relates to a group or category of consumers, from which individual consumer identities have been removed, that is not linked or reasonably linkable to any consumer or household, including via a device." Thus, although a synthetic dataset could be seen as applying to a group or category of consumers, the exclusion for aggregate data gives additional weight to the conclusion that a synthetic dataset is not covered by the CCPA definition of personal information.

Thus, because synthetic data is not "personal information" under the CCPA, it is not subject to the requirements of the CCPA. It can therefore be freely used and distributed—even sold—without restriction under the CCPA.

Issues Under HIPAA

Here we address some common questions regarding the application of HIPAA to synthetic data generation and use.

19 Some of the contract terms required under the CCPA for service providers are similar, but not identical, to the contract terms required under the GDPR for data processors. However, it is possible, and often prudent, to create terms that meet both, so that a single contract works for both CCPA and GDPR purposes.

Is the use of the original (real) dataset to generate and/or evaluate a synthetic dataset restricted or regulated under HIPAA?

HIPAA permits the use of protected health information (PHI) to create a synthetic dataset. The HIPAA Privacy Rule specifies certain uses of PHI that are permitted without the authorization of the individual and without providing the individual the opportunity to agree or object.

One such permitted use is described as follows:

> Uses and disclosures to create de-identified information. A covered entity may use protected health information to create information that is not individually identifiable health information or disclose protected health information only to a business associate for such purpose.[20]

That permitted use is reinforced in a different section of the HIPAA Privacy Rule that describes health care operations as another permitted use. Health care operations is defined to include "general administrative activities of the entity, including, but not limited to…creating de-identified health information or a limited data set."[21]

The creation of a synthetic dataset is distinct from what has traditionally been thought of as de-identification. De-identification typically involves removing, masking, or transforming direct and indirect identifiers within a record. But the resulting de-identified dataset is generally thought of as an altered version of the original dataset in which there remains some correlation between records in the original dataset and records in the de-identified dataset. By contrast, synthetic data is the creation of a completely new dataset, and while the synthetic dataset is statistically similar to the original (real) dataset, there is no direct correlation between records in the original dataset and those in the synthetic dataset.

Nevertheless, although both of these sections of the HIPAA Privacy Rule reference de-identification, both should be interpreted broadly enough to include the creation of synthetic data as a permitted use of PHI. In the first quoted section, the key phrase is "to create information that is not individually identifiable health information," which is precisely what is happening when PHI is used to create synthetic data because synthetic data is not individually identifiable data (more on that following). And in describing de-identification in that way, the HIPAA Privacy Rule strongly indicates that the concept of de-identification in HIPAA is broad enough to encompass any action that uses PHI to create a dataset that does not contain individually identifiable information.

20 HIPAA Privacy Rule § 164.502(d)(1).

21 A "limited data set" is a dataset that has had certain identifiers removed but that does not meet the HIPAA standard for fully de-identified information. See HIPAA Privacy Rule § 164.514(e)(2).

Additionally, the part of the "health care operations" definition that includes "creating de-identified health information or a limited data set" is preceded by the phrase "including, but not limited to." So, it is easy to conclude that a very similar type of operation that results in strong privacy protections for individuals is also included in that category of permitted uses. Further, with respect to both sections, given that synthetic data will almost always be even more privacy-protecting than de-identified data, there is no policy reason why these aspects of the HIPAA Privacy Rule should be interpreted narrowly or that HIPAA should treat the creation of synthetic data less favorably than the use of PHI to create a de-identified dataset.

Thus, viewing the creation of synthetic data as a permitted use of PHI under the HIPAA Privacy Rule is both a sensible and sound conclusion.

Is sharing the original dataset with a third-party service provider to generate the synthetic dataset restricted or regulated under HIPAA?

Under HIPAA, a covered entity is permitted to share PHI with another entity providing a service on behalf of that covered entity. Such a service provider is called a "business associate" of the covered entity.

There must be a contract or similar arrangement in place between a covered entity and the business associate. The contract must specify the nature of the service for which the PHI is shared, describe the permitted and required uses of protected health information by the business associate, provide assurances that the business associate will appropriately protect the privacy and security of the PHI, and meet certain other requirements.[22] Thus, a contract for the creation of synthetic data should state that the service provider may use PHI to generate and evaluate one or more synthetic datasets.[23] In addition to the contractual terms, business associates are directly subject to the HIPAA Security Rule and certain aspects of the HIPAA Privacy Rule.

Thus, a service provider that receives PHI from a HIPAA-covered entity for the purpose of creating synthetic data is likely to be considered a business associate of the covered entity. As long as there is an appropriate contract in place that meets the requirements of a business associate agreement under the HIPAA Privacy Rule, and the service provider meets its other obligations under the rule, the sharing of PHI with the service provider is allowed.

---

22 Guidance from the US Department of Health and Human Services on the required elements of a business associate agreement, as well as sample contractual language, is available at *https://oreil.ly/53Ef0*.

23 If the service provider is performing a broader range of services, and the creation of synthetic data is an inherent part of those services, the parties could argue that it is permitted even if the contract does not explicitly state that. But for the avoidance of doubt, any time a covered entity is sharing data containing PHI with a service provider to create a synthetic dataset, the parties should explicitly reference synthetic data creation in the contract as a permitted use of the PHI.

Does HIPAA regulate or otherwise affect (if at all) the resulting synthetic dataset?

Synthetic data falls outside the scope of HIPAA. HIPAA regulates "individually identifiable health information" and "protected health information." "Individually identifiable health information" is information created or received by a covered entity, relating to the physical or mental health or condition of an individual, the provision of healthcare to an individual, or the payment for the provision of healthcare to an individual, where that information either identifies the individual or for which there is a reasonable basis to believe the information can be used to identify the individual. "Protected health information" is roughly the same; it is defined as "individually identifiable health information," subject to a few minor exclusions for certain educational records and employment records.

Because synthetic data is not "real" data related to actual individuals, synthetic data does not identify any individual, nor can it reasonably be used to identify an individual. Synthetic data is therefore outside the scope of HIPAA and not subject to the requirements of the HIPAA rules. It can therefore be freely used for secondary analysis, shared for research purposes, or made publicly available without restriction.

Article 29 Working Party Opinion

The Article 29 Working Party (now the European Data Protection Board) published an influential opinion in 2014 on anonymization.[24] Although our focus here is not on anonymization, that opinion does describe European regulators' views on when information no longer becomes identifiable.[25] As well as being influential, the opinion has been critiqued on multiple dimensions.[26] Nevertheless, in the following sections we describe the criteria from this opinion for information to be non-identifiable, present our interpretation of these criteria, and explain how synthetic data would meet these criteria. At the end, we make the case that synthetic data can meet the three criteria and therefore would be considered nonpersonal information under this opinion.

The three criteria, their interpretations, and an assessment of synthetic data on each criterion are below.

24 The Article 29 Working Party is an advisory body made up of representatives from the European data protection authorities, the European Data Protection Supervisor, and the European Commission. From time to time it has published opinions interpreting and clarifying various aspects of data protection regulation.

25 Article 29 Data Protection Working Party, "Opinion 05/2014 on Anonymisation Techniques," April 2014. *https://www.pdpjournals.com/docs/88197.pdf*.

26 Khaled El Emam and Cecilia Alvarez, "A Critical Appraisal of the Article 29 Working Party Opinion 05/2014 on Data Anonymization Techniques," *International Data Privacy Law* 5, no. 1 (2015): 73–87.

Singling out

Singling out is defined as the ability to isolate some or all of the records that identify an individual in a dataset. This can be interpreted in two ways. The first is that there should be no individuals in the dataset that are also unique in the population (on the quasi-identifiers). The second is that there should not be a correct mapping between a record in the dataset and a real person.

In the case of synthetic data, there would be no unique synthetic records that map to unique real records, and hence by definition there would not be a mapping to a unique individual in the population. With respect to the second interpretation, a key premise of synthetic data is that there is no one-to-one mapping between synthetic records and individuals, and therefore this interpretation should also be met in practice.

Linkability

Linkability is the ability to link at least two records concerning the same data subject or a group of data subjects. One interpretation of this is that linkability applies to linking records that belong to the same person in the same database. This is essentially a ban on longitudinal data.[27] That interpretation has been criticized because it would have a significant negative impact on, for example, health research.

Another interpretation is that this criterion bans assigning individuals to groups, which essentially prohibits building statistical models from data (since models are based on detecting group patterns across individuals). Again, in the real world such an interpretation would halt many secondary uses of data.

Therefore, the interpretation that is generally adopted is that individuals cannot be linked across databases. This criterion is met by definition because the likelihood of successfully linking synthetic records in one database with real records in another database is going to be very low.

Inference

Inference is defined as the possibility of deducing with a high likelihood the value of an attribute from the values of a set of other attributes. One interpretation of this criterion is that it is a ban on statistics and model building, which is likely not the intent here because that would also limit the uses of aggregate/summary statistics from data involving more than one variable.

Therefore, the general interpretation is that it should not be possible to make inferences that are specific to individuals. However, inferences that pertain to groups of

27 Khaled El Emam and Cecilia Alvarez, "A Critical Appraisal of the Article 29 Working Party Opinion 05/2014 on Data Anonymization Techniques," *International Data Privacy Law* 5, no. 1 (2015): 73–87.

individuals (which is the essence of model building) would not fall within its scope. Since synthetic data does not have records that pertain to real individuals, any individual inferences would not be about specific individuals. In practice, the inferences are mostly about groups of individuals. In particular, our definition of meaningful identity disclosure would limit the information gain about specific individuals, which supports meeting this criterion.

Closing comments on the Article 29 opinion

The previous sections are a pragmatic interpretation of the three criteria in the Article 29 Working Party opinion. Synthetic data would meet these criteria in a relatively straightforward manner because it would not be matched to unique individuals, records cannot be linked across datasets, and individual-level inferences would not be possible.

Also note that some of the more general interpretations of these criteria are intended to limit risks from misuses of data and AIML models, which are best addressed through governance mechanisms and ethics reviews rather than through data transformation or generation methods.

Summary

Synthetic data is intended to protect against meaningful identity disclosure. That is, it protects against when a synthetic record is associated with a real person, and an adversary can learn something new and unusual about the target individual. Synthetic data therefore offers the promise of extracting great value from data without the privacy risk and regulatory compliance costs associated with the use of personal data or even de-identified data.

The creation of synthetic data involves the processing of a real dataset containing personal information, so the initial creation and testing of a synthetic dataset likely will fall within the scope of privacy law. But most privacy laws allow such use, subject to certain requirements such as keeping the original dataset secure and ensuring that applicable privacy notices do not preclude such use. But these are typically compliance measures that the owner of the original dataset will need to undertake in any event.

Likewise, most privacy laws allow the original dataset to be shared with third-party service providers. So data owners can provide an original dataset to a service provider that will use the data to create the synthetic data on behalf of the data owner, as long as certain data protection measures are taken. Those measures include implementing security safeguards and ensuring that an appropriate contract is in place.

And once the synthetic data is created, because it is not real data relating to real individuals, it will fall outside the scope of privacy law. It can therefore be freely used and distributed for research and other purposes.

This chapter examined three key privacy laws—Europe's GDPR, California's CCPA, and the US federal HIPAA law. Although these laws take different approaches to regulating data protection, and their details differ significantly, the conclusions regarding the creation, distribution, and use of synthetic data are similar for each. And although there can be wide variation in privacy laws across jurisdictions and sectors, they all tend to rely on similar principles and make allowances for uses of personal information that can create great social and individual benefit so long as the risks are appropriately managed. Thus, although these questions must be examined for any privacy law to which the relevant parties are subject, it is likely that the conclusions will be similar.